The impact of low-frequency genetic variation on metabolic traits of blood serum

Makenzi Nzau

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 6.6.2019

Thesis supervisor:

Prof. Jari Saramäki

Thesis advisors:

Prof. Samuli Ripatti

D.Sc. (Tech.) Taru Tukiainen

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Makenzi Nzau

Title: The impact of low-frequency genetic variation on metabolic traits of blood serum

Date: 6.6.2019 Language: English Number of pages: 5+63

Life Science Technologies

Professorship: Jari Saramäki

Supervisor: Prof. Jari Saramäki

Advisors: Prof. Samuli Ripatti, D.Sc. (Tech.) Taru Tukiainen

The effect of genetic variation on the small-molecule (metabolite) composition blood serum has hitherto been investigated with regard to common genetic variation (CV) or very-rare(Mendelian) genetic variation. However, the effect of intermediate, low-frequency variation remains less characterized.

This Thesis investigates the hypothesis, that low-frequency variation within the coding region of metabolic genes induces extreme metabolite accumulation or depletion on carrier individuals with respect to the population mean.

The sample population consists of 1159 general population individuals from the Finnish population with blood metabolite content quantified using high-performance liquid-chromatography in conjunction with mass-spectrometry.

This hypothesis is tested on selected candidate metabolites and candidate genes. The set of metabolites was obtained from prior literature of Mendelian metabolic disease, whereas the candidate genes originate from two sources. The first set originates from the mentioned disease literature, and the second set is obtained by analysis CV study results.

The population distribution of these metabolites is segregated into three groups: positive–, negative–tail and center. According to the hypothesis, the tail-groups contain an excess of carriers of above-mentioned low-frequency variation. The results of the suggest association, but not at desired significance level.

Keywords: metabolism,metabolomics,genetics,genomics, outlier, inborn errors of metabolism, loss-of-function, RVAS

Preface

The physiological state of an individual depends on a myriad of environmental factors. Examples of such factors include socioeconomic status, lifestyle choices and life-events (i.e. , history of illnesses, pregnancy). Thus, the environmental factors which shape the individual may vary vastly across individuals and time. However, genetic constitution remains invariant with respect to time and these factors. Consequently, genetic constitution provides for a plausible point of reference in the effort to tailor medical treatment to suit individual. Conversely, the state of metabolism (chemical balance within the individual) can be thought to capture an instantaneous snapshot of the individual as a biological totality.

The pursuit of the bridging of the gap between these two approaches holds great promise; to generate novel perspective on therapeutic practices and ultimately to our understanding of life as phenomenon.

From the authors perpective, the purpose of this Thesis is to investigate a particular approach to this task, hopefully contributing to the foundation of advances yet to come. Otaniemi, 6.6.2019

Contents

1 Introduction

The prior genome-wide association studies (GWAS) have been used to identify genetic variants affecting common diseases on the population scale. However, these studies have focus on common genetic variation (population frequency $\geq 5\%$). In contrast, linkage analysis has uncovered numerous causal variants of rare inherited disease in families of susceptible pedigree [1]. The exceedingly low-frequency ($\ll 1\%$) of these so called Mendelian diseases render them virtually absent from population-scale random samples [2]; these diseases are detected mainly through the clinic due to their extreme and early onset symptoms.

The intermediate frequency range (between 5% and 1%) of low-frequency genetic variation remains less investigated. The aim of this Thesis is to characterize the impact of low-frequency variation on human metabolic traits on population scale. The traits investigated in this Thesis are small-molecule (metabolite) concentrations in blood serum. The hypothesis presented in this work relies on the following postulates:

- Common variation results in small to moderate impact on metabolic traits, whereas Mendelian variation results in extreme impact (inborn error of metabolism (IEM)) — thus low-frequency variation might result in an intermediate impact between the two.

- IEM-variants induce loss-of-function (LOF) on genes facilitating metabolism — thus this Thesis examines low-frequency variants with potential to impact the function of particular genes. That is, variants in coding region (exome) of the genome.

- The causal LOF of the IEM results in cascade failure of metabolism. These failures lead to select metabolite depletion or accumulation involved key metabolites — thus the carriers of low-frequency variants might be expected to present depletion or accumulation of select metabolites.

In this Thesis, the measurement set of metabolite concentration constitute a sampling distribution. The impact of genetic variation manifests in the properties of these distributions. To accomodate this fact, the above mentioned postulates coalesce into the following qualitative hypothesis: the three regimes of genetic variation transform the distribution distinctly. The impact of common variation shifts the mean of the total distribution, whereas the impact of Mendelian variation induces bimodality onto the total distribution.

As the intermediate, this Thesis proposes as its hypothesis that low-frequency variation inflates the tails of the total distribution, as illustrated in Figure 1. More succinctly: low-frequency genetic variation within the coding region in select candidate genes should be enriched in subjects lying in the tail of selected metabolite concentration distribution. The set of candidate genes are selected by two sources : directly from prior IEM annotations and indirectly from a previous common variation association study [3].

Figure 1: The qualitative contrast between typical variant carrier and non-carrier serum metabolite concentration distributions. Common variation refers to variants with frequency $> 5\%$, ultra-rare Mendelian variants with frequency $\ll 1\%$ and low-frequency variants with $1\% - 5\%$. Common variation typically induces a shift in mean, Mendelian variation bimodality and according to our hypothesis, low-frequency variation expand the tail region.

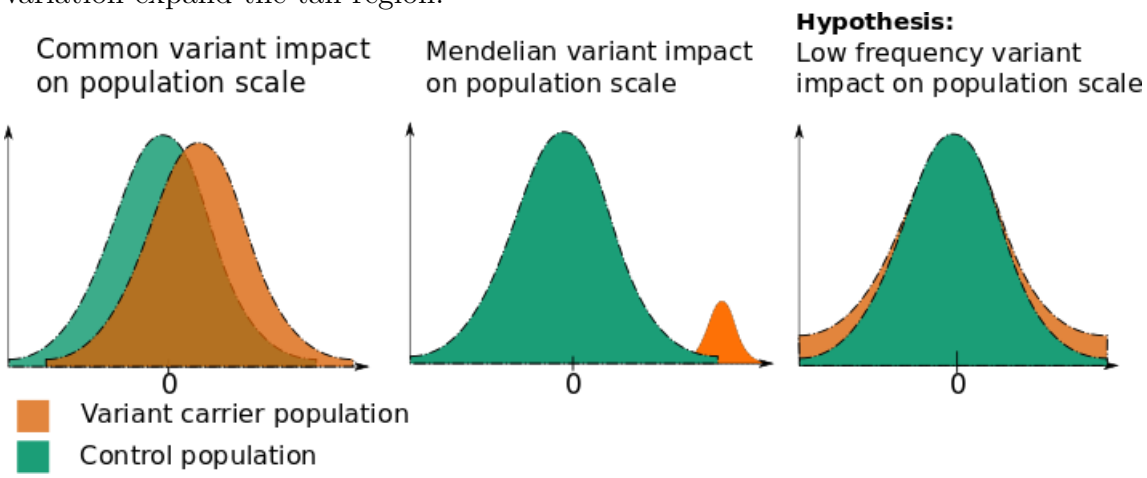
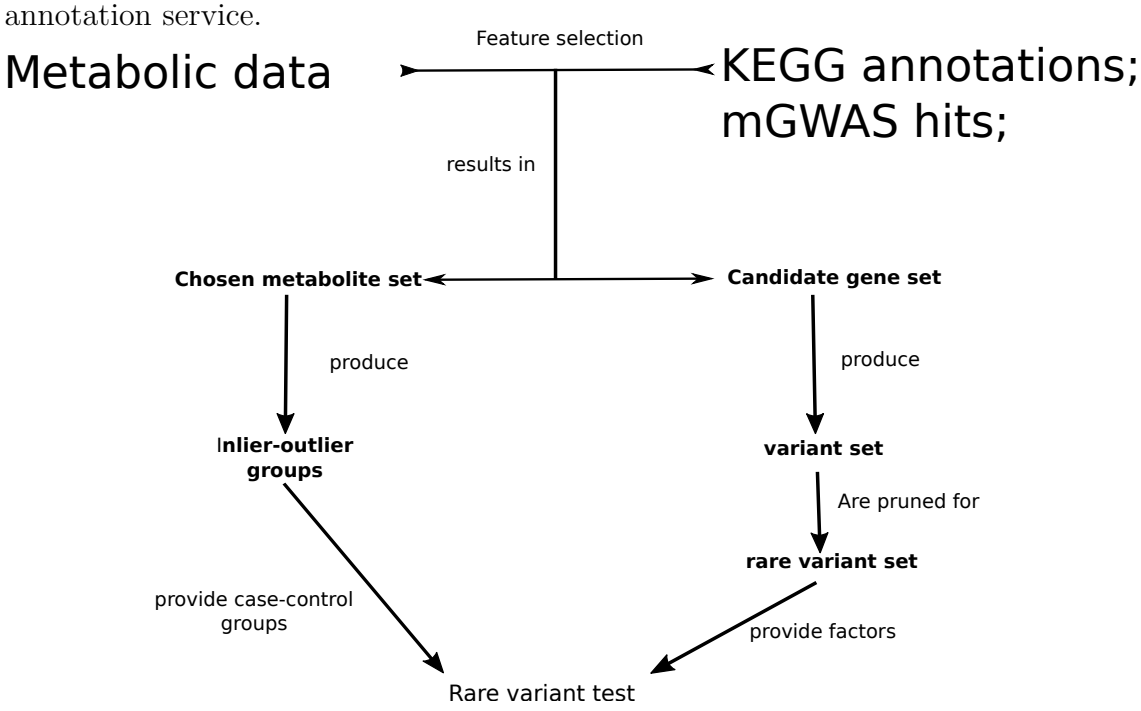


Figure 2: Coarse-grained diagram of data-flow. Features consist of candidate genes and metabolites. Metabolites are chosen based on IEM gene-metabolite pairs. The chosen metabolites are used to select candidate genes based on metabolic genome-wide association study (mGWAS) variants and IEM annotation from the KEGG [4] annotation service.



2 Background

2.1 Structure and function of the genome

The human genome consists of 23 pairs distinct units of DNA called chromosomes. These chromosomes can be further divided into two categories: the autosomal and sex chromosomes. Autosomal and sex chromosomes differ by the obligatory count inherited by a healthy individual. Figure 3 illustrates the typical female and male chromosome sets.

Each individual chromosome consists of two strands of deoxyribonucleic acid (DNA). In turn, each strand consists of an ordered sequence of four nucleotides: adenosine (A), thymine (T), guanine (G), cytosine (C). Each strand is paired with a complement strand, in which each nucleotide in replaced by their complement. A is the complement of T and vice versa, and conversely G and C are complementary. Briefly stated, each strand is coded in a quatenary code of four bases and possess a redundant complement strand. Figure 4 shows how the strand coils to form chromosomes.

The genome contains the necessary information to synthesize and regulate the synthesis of protein molecules in the cell. This information is contained in regions called genes. Approximately 1.1% of the genome consists of protein-encoding sequences and 4% key regulatory sequences and other vital non-coding sequences [5]. These protein

Figure 3: Overview of genome structure on chromosome level. The human genome consists of 44 autosomal and two sex chromosome. A: A normal male chromosome set stained for optical microscopy. The male chromosome set contains one of both X and Y chromosomes. B: a normal female chromosome set stained for optical microscopy. The female chromosome set contains two X chromosomes. This figure is a composite of Figures 5.2 and 5.3 (pg.59 , pg.60) of [5]. Modified and reprinted from "Essential Medical Genetics", by Tobias, Edward S. and Connor, Michael and Ferguson-Smith, Malcolm ; Copyright (2011) John Wiley & Sons.

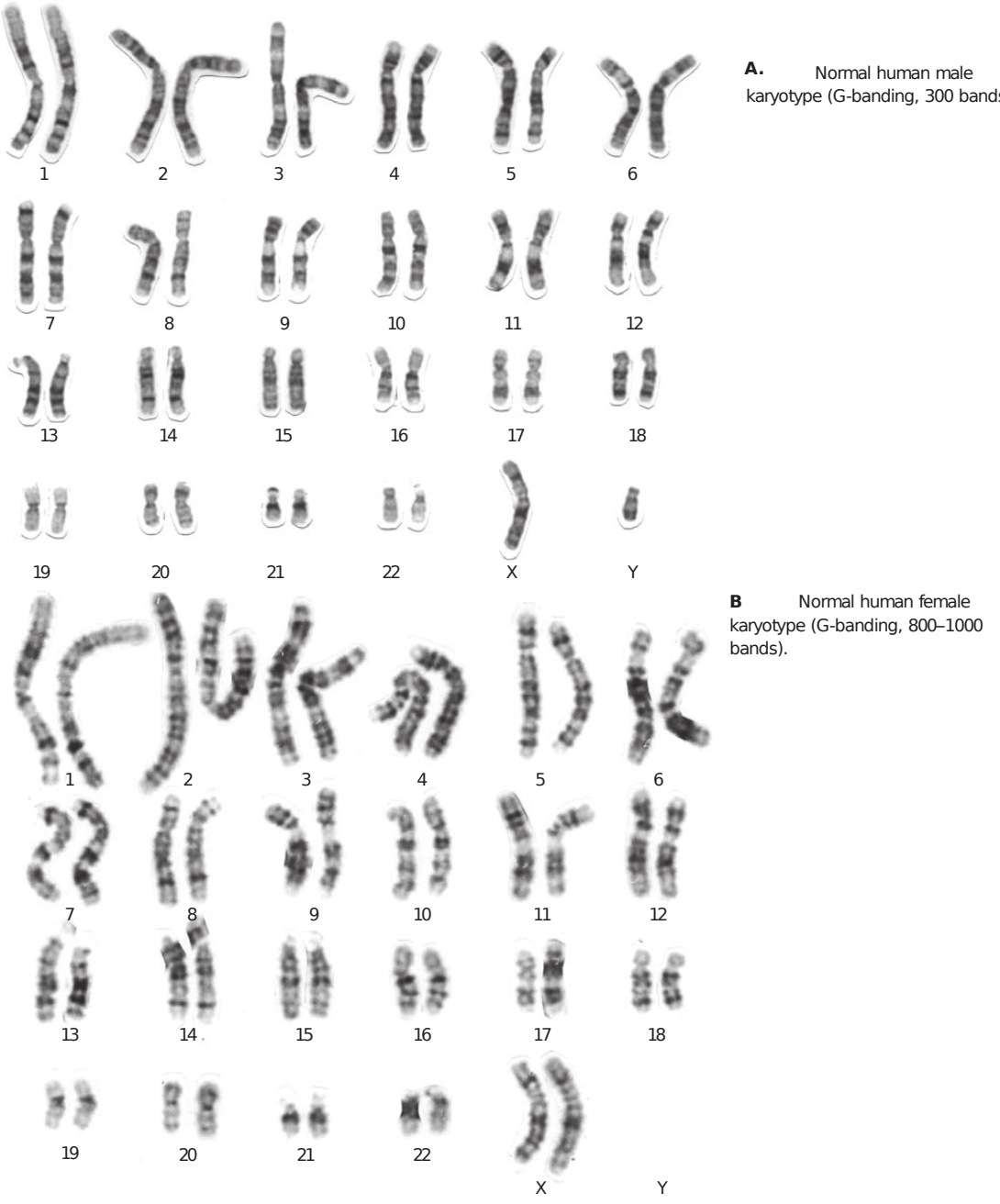


Figure 4: The composition of a chromosome in terms of DNA strand. Within the chromosome, the DNA-strand coils recursively in to the elementary fiber, which in turn is coiled into the chromatin fiber. Finally, the chromatin fiber is coiled into the chromosomal chromatid visible in the staining of 3. This figure is based on Figure 5.1 (pg.58) of [5]. Modified and reprinted from "Essential Medical Genetics", by Tobias, Edward S. and Connor, Michael and Ferguson-Smith, Malcolm ; Copyright (2011) John Wiley & Sons.

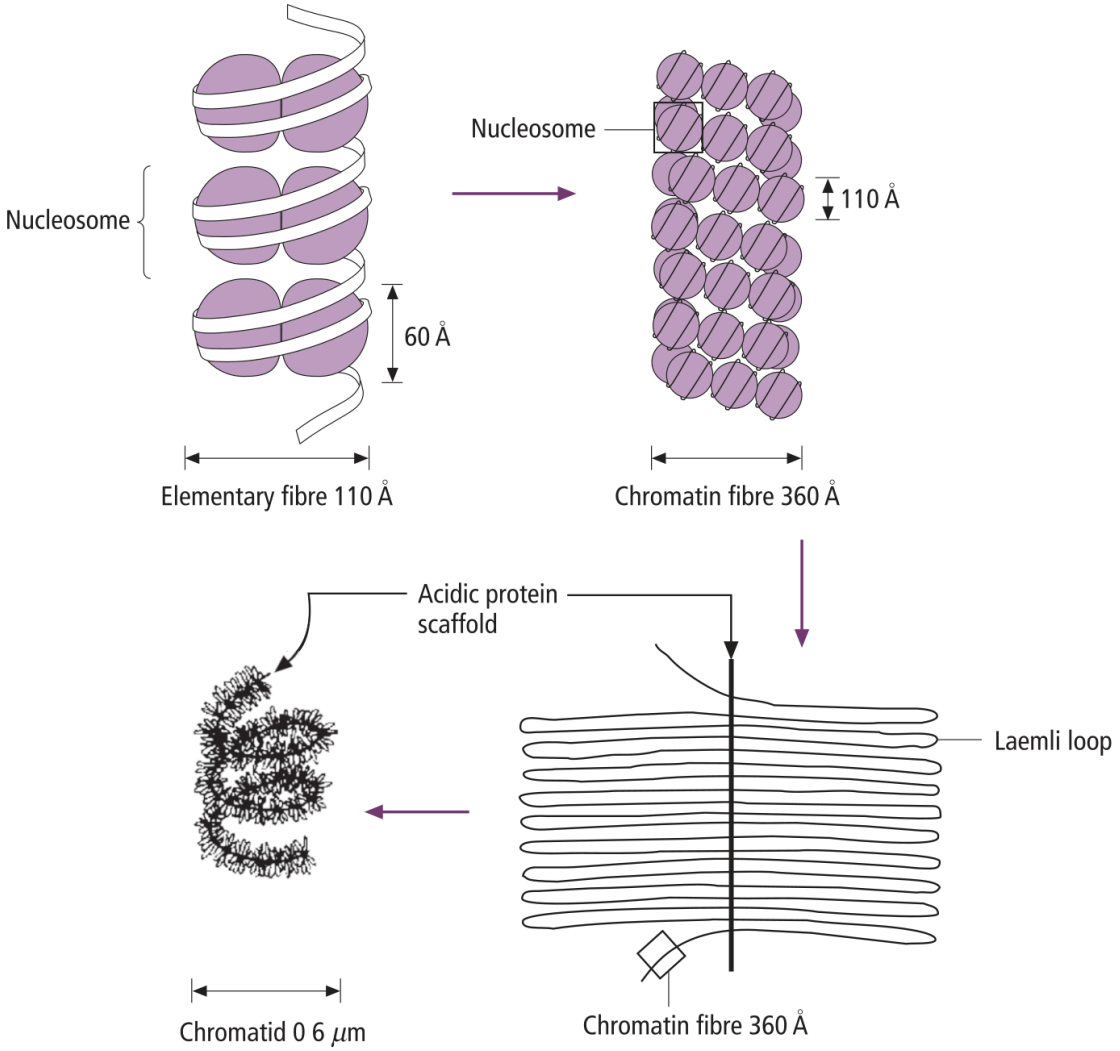
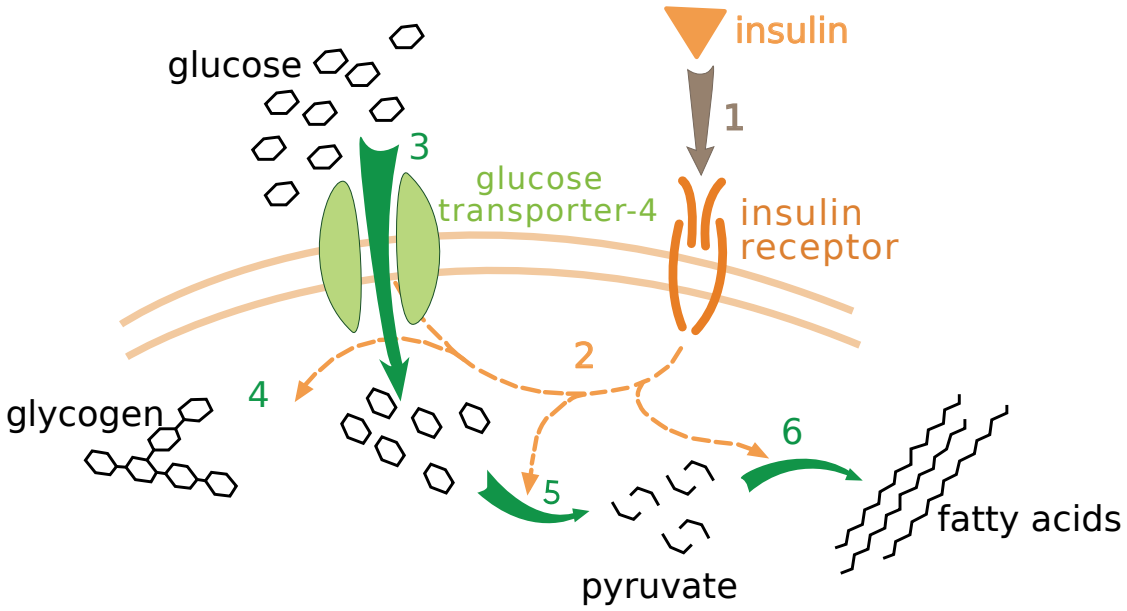


Figure 5: The protein molecules organize cell structure and metabolism. These properties enable or prevent the cell from organizing into tissues or other structures. In this figure, glucose transporter-4 and the insulin receptor organize and regulate glucose transport across cell membrane. Via wikimedia commons XcepticZP; public domain



act as the basic components of information and matter flow pathways between cells and cell compartments. As an example, Figure 5 illustrates the transport of glucose across the cell membrane to downstream chemical processes.

The function and properties of a single protein molecule is dictated by it's primary structure. From a chemical standpoint, proteins are chains of amino acids called polypeptides. These polypeptides are an ordered sequence of twenty possible amino acids. The sequence determines the three-dimesional structure which the polypeptide folds into. This three-dimensional structure in turn determines the biochemical function of the protein. Some protein consist of multiple folded chains linked by post-translational modification. Figure 6 illustrates these levels of structure via a protein called PCNA.

Figure 7 illustrates the process of information flow from the DNA strand into the synthesis of polypeptide chains. The process of deriving protein polypeptide chain from the information provided by the strand sequence is described by the central dogma of molecular biology. Briefly, the dogma dictates that the information of the DNA-molecule is transcribed into a single-stranded nucleotide sequence molecule called ribonucleic acid (RNA). Three of the nucleotides are identical; the fourth, called uracil (U), replaces thymine (T). Next, this raw transcript undergoes a process of editing called splicing. During splicing sections called introns are removed. The remaining sections (called exons) constitute the template for the polypeptide chain. This messenger RNA (mRNA) is then parsed in nucleotide triplets called codons. Each codons denotes the start of the translation, a specific amino acid or the end of translation. [5]

A gene is the region of the genome, which contains the necessary information for the transcription of a protein. Fundamentaly, genes can be divided to two subregions: regulatory and coding regions. Regulatory regions interact with the mechanisms regulating transcription rate. For example, these regions can serve as binding sites for other protein which promote or suppress transcription of the encoded protein. These regions can lie adjacent to the regulated coding region (cis-regulation) or distant to it (trans-regulation). The coding region contains the exons and the introns. [5]

2.2 Inheritance

Inheritance is the transmission of traits from parent to offspring. The genome mediates this process by storing the necessary information.

Figure 6: The protein contains three or four levels of structure. The first level (primary structure) is the amino acid residue sequence. The secondary structure is the local of folding of local section of the protein. The tertiary structure is the global folding of the protein. The fourth is the intertwining and ligation of separate polypeptide strands. The illustrated protein is PCNA. Via Wikimedia Commons by Thomas Schafee (Evolution and evolvability); cc BY-4.0 license; [6]

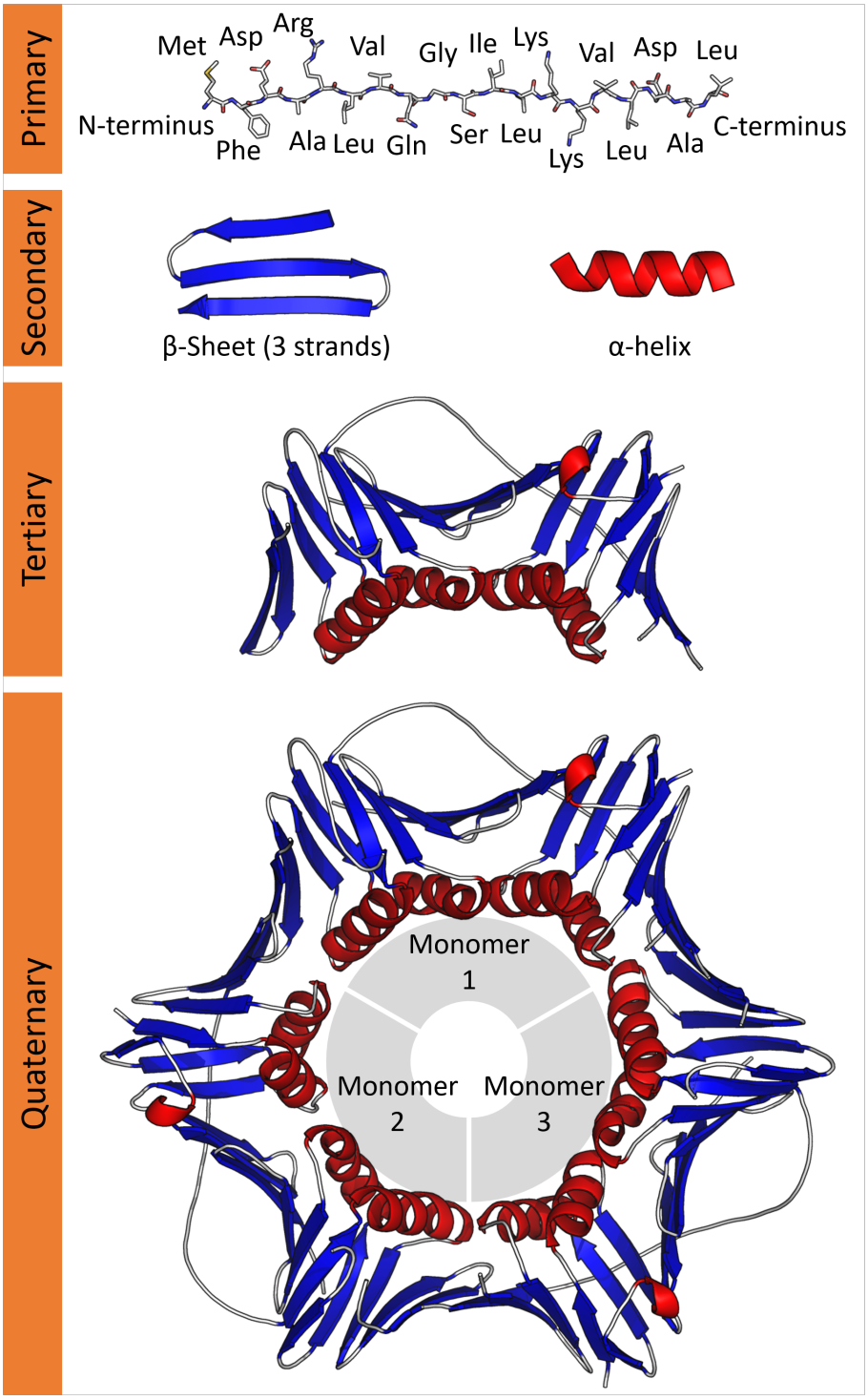
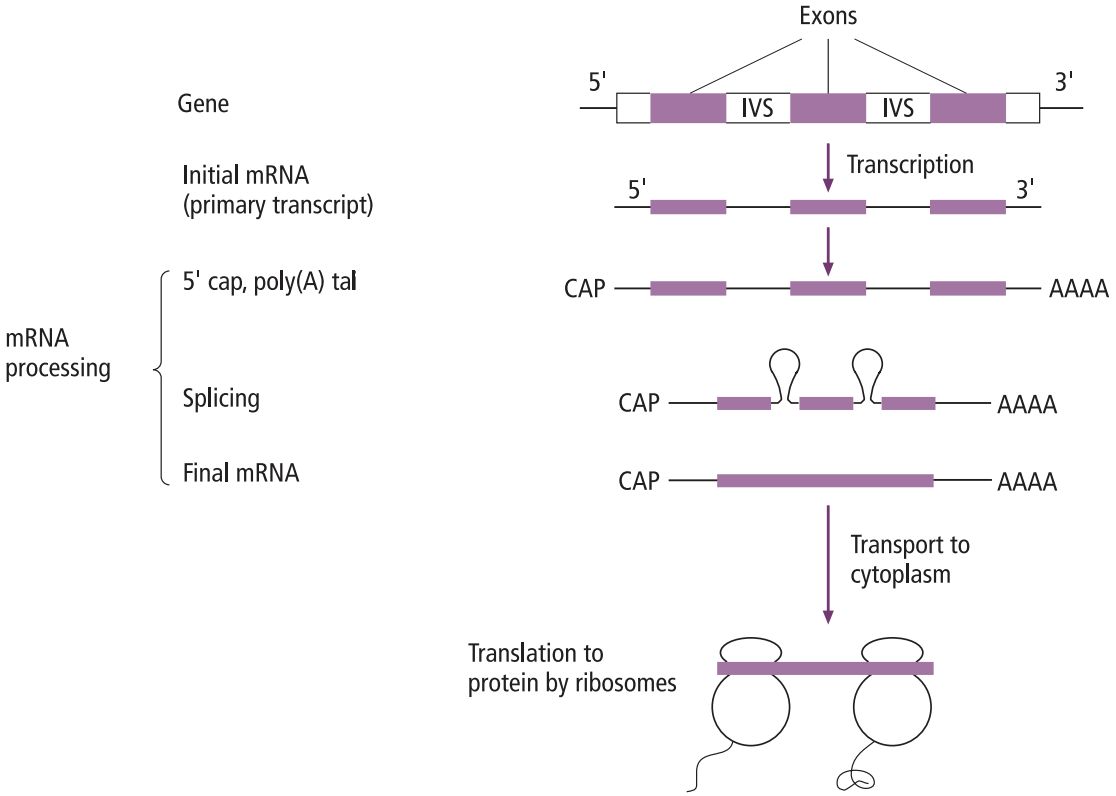


Figure 7: The production process of protein polypeptide chain from genetic DNA information. The DNA sequence is transcribed into mRNA; the primary mRNA is purged of introns, leaving only concatenated exon sequences. The final mRNA is translated into the polypeptide chain. Modified and reprinted from "Essential Medical Genetics", by Tobias, Edward S. and Connor, Michael and Ferguson-Smith, Malcolm ; Copyright (2011) John Wiley & Sons.



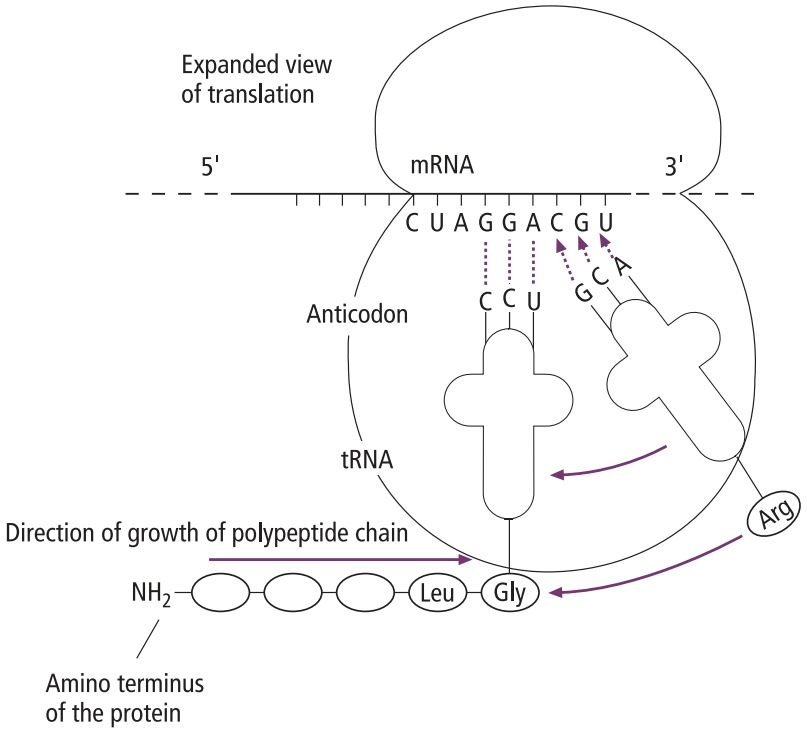


Figure 8: Mendelian inheritance with diploid organisms: 1. Definition of diploidy: each individual carries two copies of chromosomes. The blue chromosomes confer the blue phenotype and the red chromosomes the red phenotype; 2. The formation of sex-cells entails fission of the diploid genome into single-copy haploid genomes; 3. As result of mating or crossing, the haploid sex-cells fuse resulting in diploid offspring; 4. Dominant and recessive trait: the red trait dominates the blue, recessive trait. That is, all red-blue heterozygotes present the red trait. Via Wikimedia Commons (modified) Asychterz18;(CC BY-SA 4.0 license); [7]

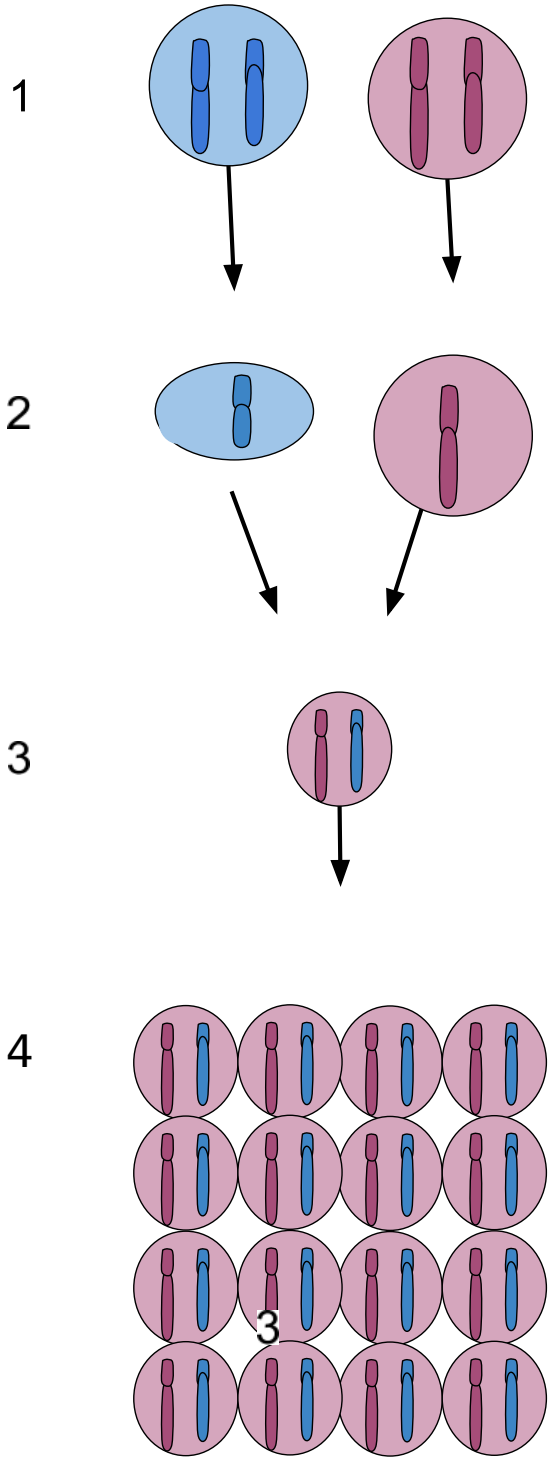


Figure 9: The principle of homologous recombination: The maternal (M) and paternal (F) chromosomes recombine by exchanging sections. The result is twohaploid genomes of two sex-cells. Via wikimedia commons David Eccles (Gringer); cc-by-2.5 license [8]

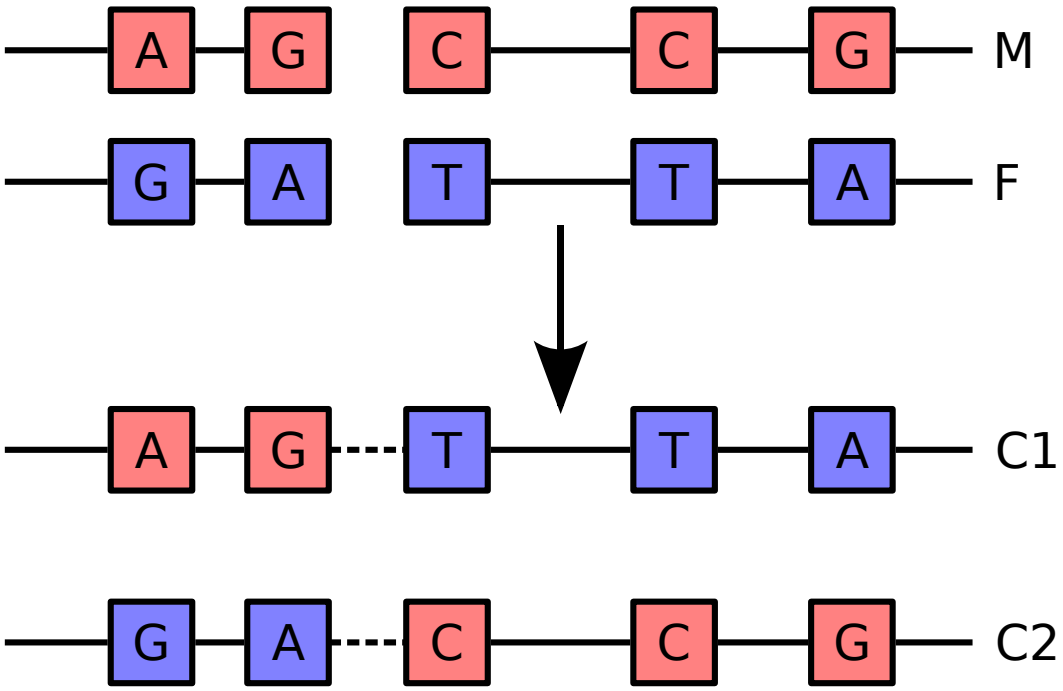
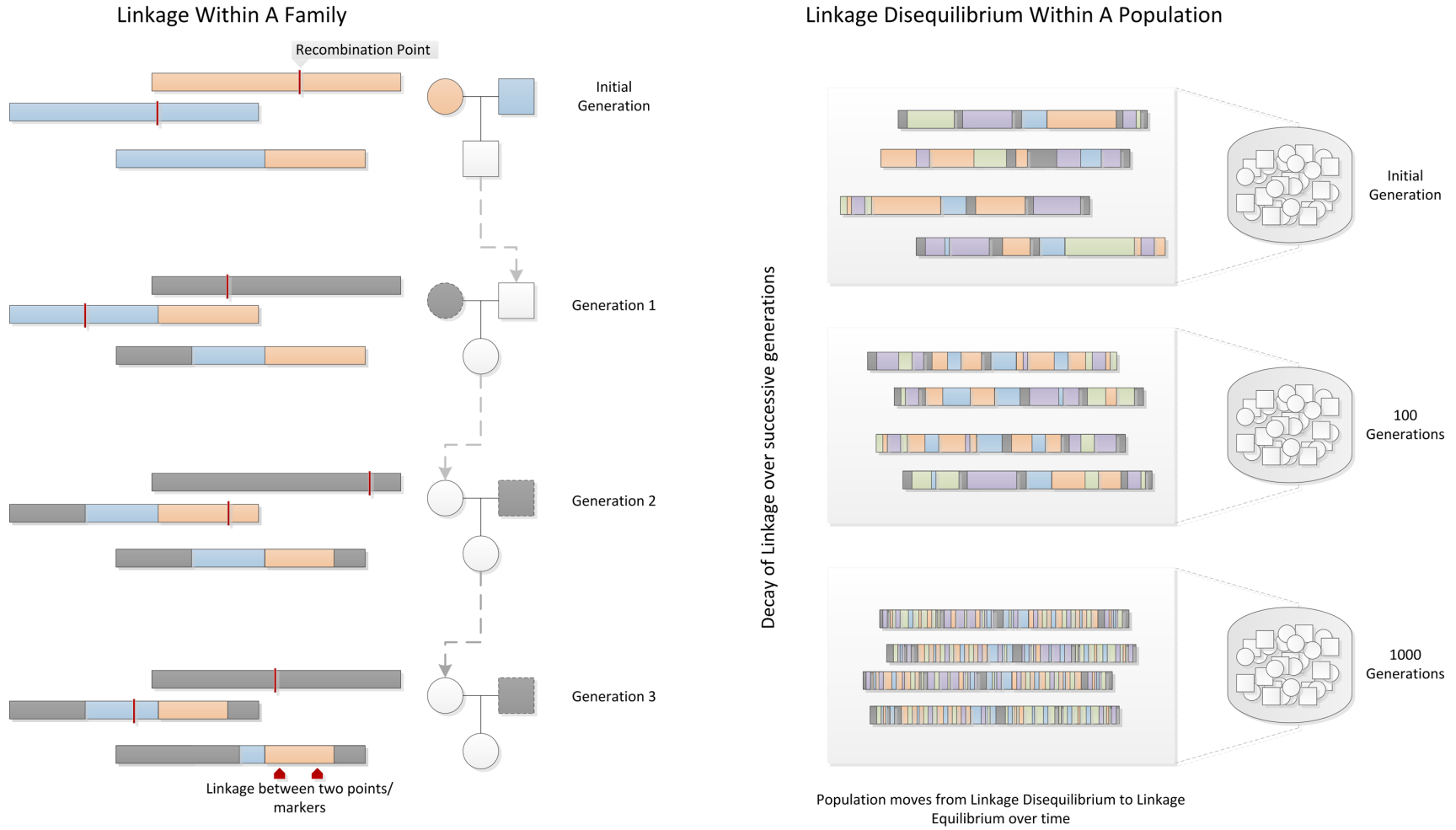


Figure 10: Origin of linkage and LD. Linkage within a family: linkage is the result of approximate conservation of crossover sites. T. Typical segment (or haplotype) lengths are large enough to conserve linkage over few generations. This results in correlation between alleles which share haplotype; LD within a population: Linkage can observed on population scale. Alleles correlate over significant distances. However, due to randomness in recombination, the correlations decay over generations leading to an uncorrelated equilibrium. Thus, linkage disequilibrium tends towards linkage equilibrium. Figure by William S. Bush, Jason H. Moore via Wikimedia commons; cc-by-3 license [9]



Even though the detailed molecular basis of inheritance is complex, it can be conceptualized using two events: the transmission of the genome and the development of traits from said genome.

The individual obtains one chromosome per chromosome pair from each parent. This chromosomes is derived via recombination of the respective parents corresponding chromosome pair. Thus each parent contributes a recombined half-genome (haploid) to the offspring (figure 8).

The generation of haploid genomes from diploid genomes is called meiosis. During meiosis, the parental chromosomes pair by chromosome type. Then, the both chromosomes exchange segments by process called homologous recombination as illustrated in Figure 9. Finally, the recombined chromosomes segregate, forming two types haploid genomes.

The form of this process leads to two noteworthy consequences. First, only one allele can be inherited from two ancestral alleles of the parent. Second, the randomness inherent to the selection of end and start loci of recombined segments implies decoupling of loci within the ancestral chromosome. The decoupling has been observed between chromosomally distant locus , but with lower frequency between adjacent locus. This phenomenon is called chromosomal linkage (CL) and it results in correlation between alleles in adjacent loci. This phenomenon vanishes in on population scale on the long-term, but population bottlenecks can sustain linkage between alleles. This phenomenon is called LD and it has profound implications on population genetics. Figure 10 illustrates the origin and evolution of LD across generations in the population.

The development of traits from the genome depends on the examined trait. The most elementary (or Mendelian) traits vary due to variation in individual loci. That is, hetero- or homozygosity with respect to the causal locus determines the presented trait. Figure 8 illustrates an example of this. However, common diseases such as coronary artery disease (CAD) or type 2 diabetes (T2D) or continuously quantitative traits (height, BMI) are not accurately explained by the Mendelian mode of inheritance. In contrast to Mendelian traits, these traits are often influenced by the environment and depend on a myriad of loci.

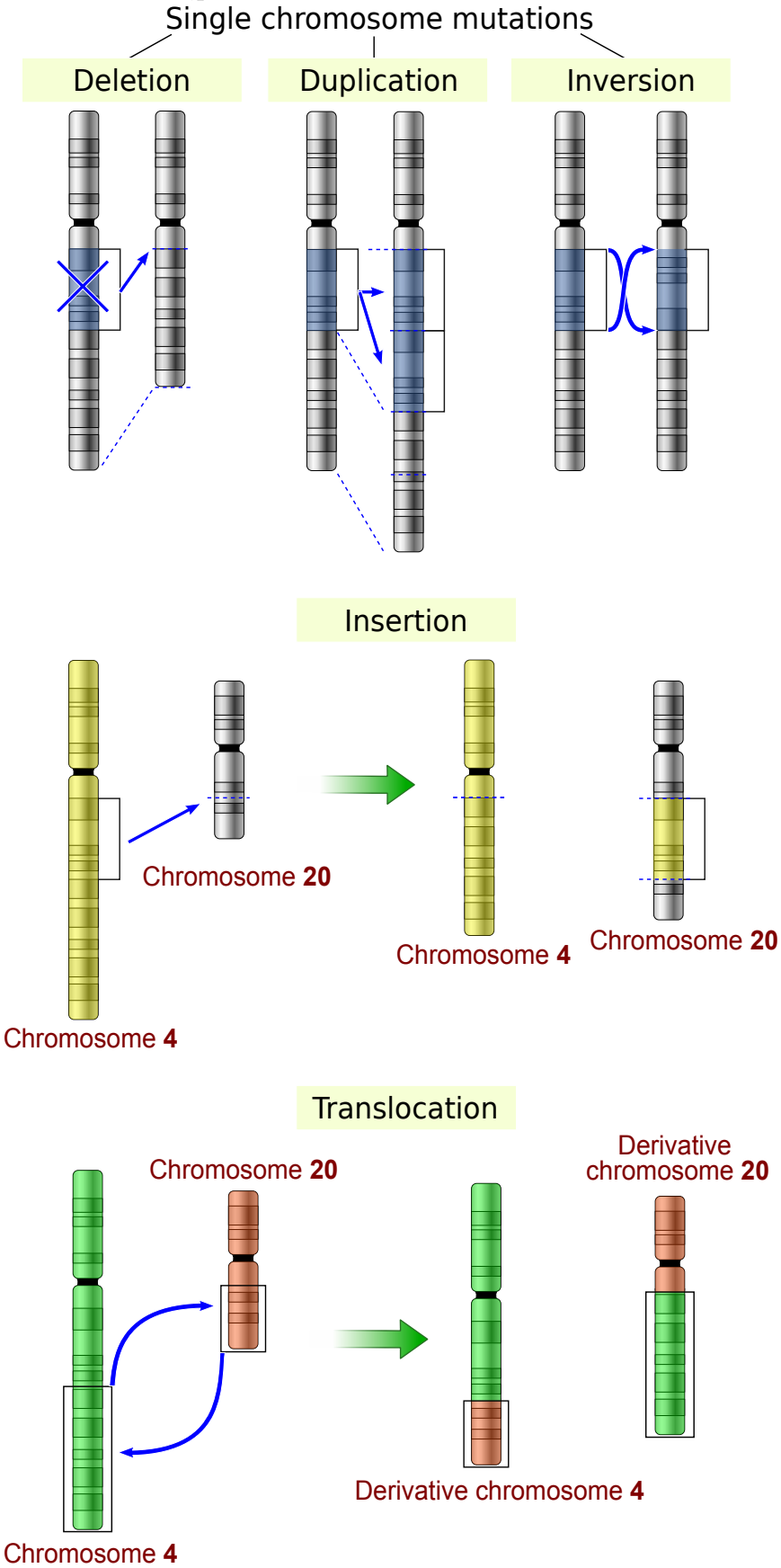
The most elementary model used for these traits is an additive model within and between loci. That is, each allele affects the trait independently and the sum of effects totals the genetic variance of the trait.

3 Genetic variation

The genetic sequence varies from individual to individual. This variance may lead in individual differences in traits such as height or disease risk [10, 11]. These variations are either inherited (the mixing properties of meiosis) or produced *de novo* by DNA damaging agents (UV radiation, mutagenic chemicals) or by errors in DNA replication. [5]

Genetic variation can be categorized in several ways: by structure, effect or frequency in the population of interest. Structure-wise, variants can be divided

Figure 11: Illustration of multiple types of length mutation at chromosome scale. Via wikimedia commons ;public domain



into two categories: point mutations (or single nucleotide polymorphism (SNP)) and length mutations. Point mutations involve a substitution of a single nucleotide, whereas variants resulting in insertions or deletions are called length mutations [5]. Insertions or deletions can range from length of several nucleotides to entire sections of the chromosome (chromosomal translocation). Figure 11 illustrates the spectrum of length mutations discussed above on the scale of the chromosome. In the scope of this Thesis, all variants considered are SNPs. In this case, the functional impact of a variant depends on the site of the variant. Specifically, whether the variant lies within a protein-coding or a regulatory region of a gene.

At the most elementary, variants substitute a codon for another. If a substitution of synonymous codons occurs, the sequence remains unaltered. These variants are called silent mutations. Variants causing non-synonymous substitutions are called missense mutation. variant resulting in premature stop-codons are called non-sense mutation [12]. More complex alterations can be caused by splice-site mutations. Splice-site mutationsare variants on the intro-exon boundary, which alter splicing process. This can lead to inclusion of entire introns or exclusion of exons.

Thus in principles, exonic variants may impact the host protein significantly. Non-sense mutations either produce truncated or no protein.Missense mutations in turn may produce protein, but with typically reduced or no activity. The consequences of splice-site mutations are more diverse, and hence constitute an independent class [13]. The typical consequence of these mutations is LOF in the host protein. However, gain-of-function mutations have been observed as well [5]

Most common variants impact the individual at best moderately. Exceptions have been discovered, such as several variants harboured by the APOE gene, which significantly increase the risk of late-onset Alzheimer's disease in European populations [14]. By contrast, known rare variants are likely harm to individual. For example, causal variants of cystic fibrosis are SNP which are rare among the general population but enriched in affected families. Other examples of such variants underlie Crohn's and Huntington's disease [15].

Genetic variation has been studied primarily on two scales: family and population. Family-scale studies investigate the effect of variation by comparing trait variation between and within sets of related individuals, whereas in population-scale studies examine general effects independent of descent. Family-scale studies have discovered the genetic basis of a myriad of rare inheritable diseases, but fail to explain the heritability of common diseases. This observation serves as the basis of tthe common-disease-common-variant hypothesis. This hypothesis states that single high-impact variants explain rare disease risk, whereas multiple minor or moderate impact variants explain common disease risk. Given the large number and minute impact of the sought variants, large sample sizes are necessary for statistical power. Thus, population-scale studies arise as solution.

3.1 Genome Wide Association Study

The effect of common variation on a phenotype within a population is investigated by means of genome-wide association study. The common variation consists of

selected single nucleotide polymorphism loci; the population is then sampled for a representative sample. Then, the these sampled individuals are genotyped and measured for the chosen phenotype. The task of GWAS is to determine for each locus, whether carriers of similar alleles agree by phenotype.

Two aspects define the statistical framework of GWAS: the study design and genotype model. Study design depends largely on the phenotype. Dichotomous phenotypes can be examined with a case-control design, while continuous traits are typically examined quantitatively. The genotype model dictates the grouping of genotype. For example, a two-allele (major allele A, minor a) dominant model groups the genotypes aA, AA, Aa together and AA as the other group. [15]

The typical quantitative trait locus analysis (QTL) study has a continuous phenotype with an allelic additive model. The allelic additive model groups genotypes by allele count. For a two allele locus, the classes are 0,1,2 copies of the minor allele. Each copy is assumed affect the quantitative phenotype identically. The effect can be statistically measured and tested with generalized linear models such as ANOVA [15].

4 Metabolism

Metabolism refers to regulated transformation and transport of chemical species (metabolites) within the body or across its boundaries. Metabolites are typically transformed in chemical reactions catalyzed by protein called enzymes. Most reactions are catalyzed by a single, specific enzyme, which without the reaction would not proceed. Thus, the activity and the concentration of the key enzyme regulates the balance between the substrate and the product concentrations.

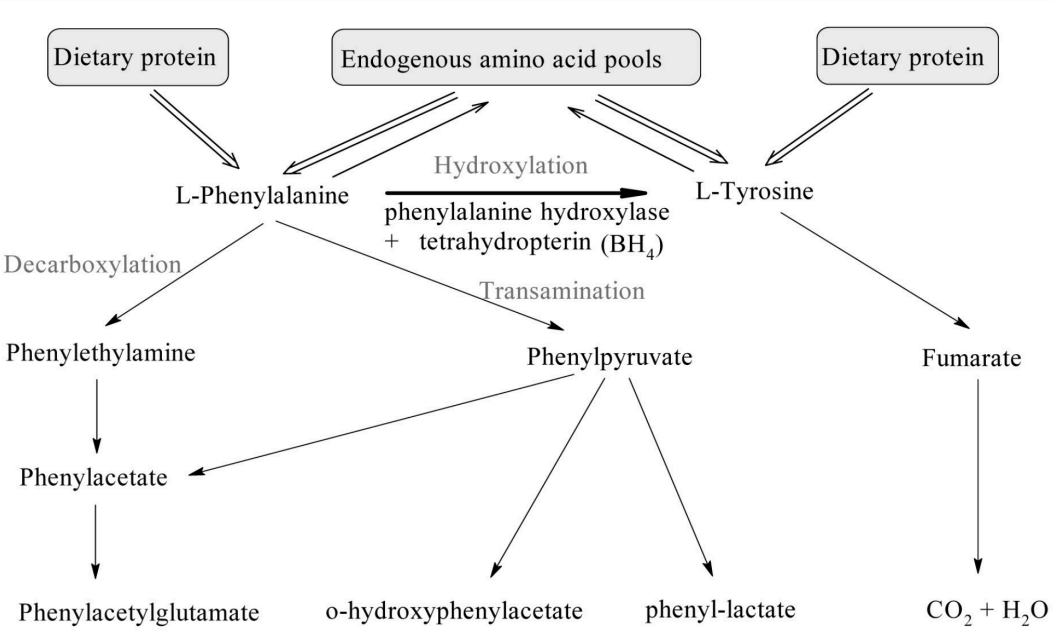
These reactions can assemble into linear, cyclic or branching sequences called pathways. These pathways are typically separated by physical compartments. These compartments can be cellular organelles such mitochondrion or tissues. Transporter protein transport chemical species across the barriers between these compartments. Transporters may function uni- or bidirectionally and vary in specificity.

In the context of this Thesis, all metabolites are small-molecules, as opposed to large macromolecules such as protein peptide chains. Figure 12 demonstrates phenylalanine metabolism. The reaction from phenylalanine to tyrosine is catalyzed by the enzyme phenylalaninehydroxylase (PAH). The disruption of the gene encoding for PAH leads to the IEM phenylketonuria (PKU) [45].

4.1 Metabolomics

Metabolomics is the high-throughput study of small-molecule compounds and their biochemistry in biofluid or tissue. Measurements may provide up to hundreds of small-molecule metabolites concentrations simultaneously. These metabolites can be endogenous compounds such as amino acids, carbohydrates, nucleotides or lipids or exogenous compounds such as pharmaceuticals, food additives or pesticides. Thus, metabolomic methods are ideal for measuring the state of complete biochemical substructures, such as metabolic pathways [16].

Figure 12: Amino acid metabolism responsible for the IEM phenylketonuria (PKU). The key metabolites phenylalanine and tyrosine are obtained through dietary intake and amino acid recycling. They further transform into metabolites downstream. The enzyme phenylalaninehydroxylase (PAH) catalyzes the conversion of its substrate (phenylalanine) to the product, tyrosine. Disruption of this conversion results in PKU. Several Mendelian variants are known to induce LOF on PAH. Figure 1 of Williams, R. A., Mamotte, C. D., & Burnett, J. R. (2008); Copyright (2005) The Australasian Association of Clinical Biochemists Inc.



Spectroscopic techniques provide the foundation for metabolomic methods. The two most popular families of techniques are mass spectrometry (MS) and hydrogen nuclear magnetic resonance (H-NMR), detect compounds in atto- and micromolar resolution respectively [17]. Since analysis require little biofluid or tissue, these techniques scale to epidemiological sample sizes. Additionally, techniques from both families can be used to quantify previously unknown metabolites. Metabolomic studies which also quantify unknown metabolites are called untargeted studies. Conversely, Metabolomic studies with a priori chosen metabolites of interest are called targeted studies.

Although the two families share numerous attributes, they differ in several key aspects. Briefly, H-NMR-based analyses involve minimal sample handling, are generally non-destructive and produce easily quantifiable results. These benefits come at the cost of specificity and sensitivity. Namely, the lower spectral resolution of H-NMR based techniques leads to reduced distiguishability between similar compounds. Likewise, the sensitivity to compound quantity is orders of magnitude lower than that of MS-based. Conversely, MS-based techniques gain sensitivity and precision at the cost of introducing complexity into the analysis. Factors, such as compound ionizability and ion suppression, must be accounted for gain precision. Furthermore, MS-based analyses often feature chromatographic separation prior to spectral measurement. This increases peak separability but at the risk of introducing nuisance factors [17].

There are numerous biofluids and tissues of interest. Typically studied biofluids include blood, saliva and urine but even cerebrospinal fluid has been used [18]. The primarily metabolomic studies have focused on blood and urine. This is due to the fact that these fluids represent the global state of metabolism.

5 Genomic analysis of metabolomic phenotypes

The above mentioned techniques enable metabolomic studies with thousands of samples, rendering association studies to genetic variation feasible on epidemiological scale. As all descriptions of physiological processes are in theory metabolic, metabolomics does not impose strict boundary condition to study structure. Instead, the phenomenon of interest and it's genetic architecture provide the boundaries for study structure. As a result, the variety of developments in study structure and composition have taken place.

Three aspects outline the dimensions of development: increase in study population size and diversity, increase in genotyping depth and increase in phenotype structure depth. Initial studies had relatively low sample size sampled from a single cohort, genotyped with low resolution and little or no structure in metabolic phenotype. The subsequent studies have developed with respect to one or more of these aspects.

5.1 Cohort size and diversity

The first studies were genome-wide association study on measurements of metabolite content of blood. The study structure standard QTL mapping, hence the term

metabolic quantitative locus (mQTL) for alleles influencing metabolite content [19, 20, 21].

Gieger *et al.* [19] noted that using concentration ratios instead of individual concentration values produced a sharp increase in significance (p-value) of discovered associations. The proposed interpretation for the phenomenon is that a certain subset of these ratios represent the substrate-product equilibrium in enzymatic reactions (substrate-product relation illustrated in Figure 12). Gieger *et al.* [19] present variants discovered on the FADS1 and LIPC genes. Similar studies have been performed with larger sample-sizes and numerous cohorts [22, 3, 23]

5.2 Phenotype structure depth

Prior biochemical or clinical knowledge can be exploited in a myriad of ways. The subsequent mGWAS-derived studies vary in their approaches to exploitation. These approaches can be laid on a spectrum of *supervisedness*. *Supervised* approaches impose specific a priori biochemical pathway or clinical onto the metabolite or genotype data. In contrast, *unsupervised* are *impose* an abstract, general structure upon the data (e.g. pathway or gene-gene interaction network).

In the study conducted Hartiala *et al.* [24], the role of a specific biochemical pathways in CAD risk is investigated. Hartiala *et al.* investigate the effect of betaine metabolism on CAD risk mediated by choline metabolism. Specifically, the study focused on two competing pathways of choline metabolism: the TMA pathway and the betaine pathway. The activity of these pathways have been implicated in an increase and reduction of CAD risk respectively. Based on this premise, they first performed a genome-wide association study on plasma betaine levels. The discoveries comprised the set of variants of interest. This set was then tested for association with metabolites of interest; that is, a subset of the intermediate metabolites of the betaine pathway. Finally, the immediate and metabolite-mediated difference in CVD risk was evaluated and compared for each variant.

In the unsupervised end of the spectrum lies the mGWAS by Mittelstrass *et al.* [25]. Instead of leveraging *a priori* pathway knowledge, they construct a gaussian graphical machine (GGM) onto the set of metabolites. This GGM models the correlation structure of these metabolites as graph with metabolites as nodes and significant correlations as edges. The GGM then provides a context which in sexual dimorphisms in serum metabolite content and subsequent sex-stratified mGWAS results are analyzed. Specifically, effect of sex on serum metabolite concentration was quantified with linear regression. The inferred regression coefficients placed on the GGM network as vertex weights. Then, the community structure of connected, similarly weighted vertices were analyzed. GGM-based approaches have been extended and applied in other contexts [26].

5.3 Genotype resolution depth

Advances in genotyping resolution has been leveraged by mGWAS-based studies. Particulary exome sequencing has been used in metabolic genetic association stud-

ies [27, 28, 29, 30]. Exome sequencing provides two key advantages to general genome-wide association study: power to detect rare variation and context for discovered variants. Since exomes are gene-encoding regions, the relevance of discovered variants can be more readily evaluated. On the functional level, the change in enzyme or transporter activity can be readily interpreted in clinical or experimental settings. On the structural level, the impact of variant can be evaluated by its consequence on the produced polypeptide chain.

Yu *et al.* investigate the effect of variants predicted to cause LOF in the host gene. Such variants were defined as SNPs and small indels resulting in truncated protein or non-viable transcripts. More specifically, they selected variants causing premature stop-codons, frameshift or splice-site disruption. Then, they quantified variant effect by regressing metabolite levels on allele count par with standard mQTL analysis.

6 Inborn errors of metabolism

IEM are defined as inherited metabolic diseases caused by ultra-rare (frequency in population $\ll \%1$) variants. The hallmark of these diseases is extreme accumulation or depletion of signature metabolites.

As with other Mendelian diseases, due to their extremely low-frequency, these diseases have been typically studied in families enriched with the causative variants. [1]

Currently, there are numerous known IEM with diverse clinical and biochemical manifestation. Thus multiple types of taxonomy have been formulated. Given the molecular focus of this Thesis, the following taxonomy presented by Salmi *et al.* [31] fits this Thesis well.

IEM are typically caused by the LOF of enzymes or enzyme co-factors facilitating biochemical reactions, or metabolite transport protein between compartments (illustrated in 12). The impact of the LOF on the target reaction and pathway can thus be described followingly,

1. Accumulation of enzyme substrate,
2. Depletion of enzyme product,
3. Accumulation of typically minor metabolites,
4. Secondary metabolic consequences ,e.g, acidosis or ketosis.

The first two features result directly from failure of the target reaction, whereas the latter can be ascribed to systemic disruption of metabolism on pathway scale. The first three impacts generate the signature mentioned above.

The prior research of IEM therefore shows that these signature metabolites can exhibit extreme variation caused by genetic variation. Thus, the set of signature metabolitesconstitute an ideal candidate set of metabolites. Furthermore, the causal genes constitute plausible set of candidate genes.

The annotation service KEGG [4] provided IEM annotations for this Thesis. These annotations map IEM genes to associated metabolites.

The dataset contains 9 IEM metabolites corresponding to 9 IEM 3. Furthermore, these annotations provide a list of candidate genes per IEM 4.

7 Methods and materials

7.1 SNP causality - candidate genes

In addition to the IEM gene-set, this Thesis also considers genes implicated by the prior mGWAS of Shin *et al.* [3].

Establishing the mechanism of a mGWAS discovered SNP on the trait of interest requires two steps. First, the discovered SNP may present a false positive discovery due to LD (see Subsection 2.2). That is, the discovered variant might occur merely in correlation with the true causative variant [32]. These tag SNPs can be filtered using LD-pruning. The process of LD-pruning examines the strand in windowed segments. These windows are centered around each candidate variant and are fixed length-wise. Windows containing many candidate variants are likely to contain only one causal single nucleotide polymorphism and multiple tag single nucleotide polymorphisms. The most significantly associated variant is selected by p-value as the causal single nucleotide polymorphism and the remaining are discarded. These causal single nucleotide polymorphism serve as genetic landmarks to identify the structure of the pathways or mechanism of interest.

To establish the mechanism of effect, the causal SNP is mapped onto set of plausibly affected genes. Genes located near causal single nucleotide polymorphisms map onto the candidate set. The simplest mapping is windowed selection of nearby genes. However, this task is untrivial as many discovered SNPs lie in intergenic regions. In other words, the shortest distance from single nucleotide polymorphism to gene can be extremely large and vary vastly from genetic region to region. This can somewhat accounted for by appropriately large window-size. [33]

A large window in the dense region of the genome results in disproportionate variation in the number of candidate genes per single nucleotide polymorphism. Limiting the number of candidate genes selected per single nucleotide polymorphism evens the proportions.

In this Thesis, the mGWAS discoveries of Shin *et al.* [34] for the selected candidate metabolites provide the candidate single nucleotide polymorphism set. The following process selects the mGWAS candidate gene-set per metabolite.

Preprocessing: Prior to LD-pruning, all discoveries with $p \geq 10^{-5}$ were discarded.

LD-pruning: The set of remaining discoveries were LD-pruned with window-length of 500kb[1] resulting in a set of causal single nucleotide polymorphisms

Gene selection: For each causal single nucleotide polymorphism, genes within the 15kb window were selected.

[1] kb short for kilobase

Gene count normalization: Finally, the ten closest genes to each window center (that is, causal single nucleotide polymorphism) were selected to the final set.

7.2 Rare variant analysis and extreme phenotype sampling

Rare variant analysis presents a challenge to the statistical methods used in common variant association studies: the lack of statistical power. This challenge can be mitigated by increased sample sizes or large effect sizes of discovered variants. However, an alternative to these measures is using specialized statistical methods and study designs [35].

This Thesis adapts two methods from the above mentioned publication: extreme phenotype sampling (EPS) and region-based testing.

The first method adapted into this Thesis is EPS. EPS involves sampling the tails of the phenotype distribution. Typically this distribution is assumed or deliberately transformed to normal distribution. As a result, definition of the tail region becomes intuitively and rigorously defined. This definition will be further explored in **??**.

Region-based test aggregate adjacent variants into a combined region. The region-based hypothesizes, that this region comprises a functional unit. Therefore, the number of minor alleles carried by an individual in this region determines their phenotype. In other words, the individual effects are compounded into a single quantity. The effect of this total quantity is then tested for association with the phenotype. In the context of this Thesis, these regions are candidate genes. As mentioned before, this Thesis aims to investigate plausible LOF variants in candidate genes. Consequently, this leads to two criteria for variant selection:

- The variant site should lie within or near an exon and preferably encode a non-synonymous mutation mutation.
- The variants should present a minor allele frequency (MAF) less than or equal to 5% in the finnish population according ExAc

Given the division of samples into tail and center groups and the aggregation of variants into compound regions, the objective of this Thesis can be formulated as follows: to detect enrichment in the number mutant alleles by region within the tail groups in contrast to the center group. Lee *et al.* list multiple methods designed to achieve this task.

This Thesis employs the following design: Instead of compounding the mutational burden over individuals, it is compounded over the group. That is, the total frequency of minor alleles within a tail population against the central population. The association is tested with the χ^2 test.

The χ^2-test measures the association between two categorical variables. The domain of both variables consist of independent and mutually exclusive categories. Thus an association between two variables is measured by the number cross-occurrences between categories of these variables. McHugh *et al.* have elaborated this method further in their 2013 publication [36]. The population variable consists of three mutually exclusive categories: positive–tail, negative–tail and inlier. The genotype

variable per metabolite per gene consists of two categories: carrier and non-carrier. The mGWAS and IEM gene-sets are tested independently of each other.

Formulated in terms of χ^2 quantities: we wish to test whether the observed count of variants in either tail O_+, O_- differ significantly from the expected count $E[+], E[-]$. The expected count is derived from the frequency of carriers in the inlier population. These are compared via the test statistic $Q_+, Q-$.

Stated in terms of null– and alternative, we have the following hypotheses:

Null-hypothesis (1) $p_+ = p_{inlier}$,

Null-hypothesis (2) $p_- = p_{inlier}$,

Alternative hypothesis: $p_+ \neq p_{inlier}$ or $p_- \neq p_{inlier}$.

Here, $p_\pm$ refers to the proportion of variants in the positive (or negative) tail group and p_{inlier} to the proportion of carriers within inlier group.

In terms of the test statistic $Q_{i,\pm}$, according to the null-hypothesis we have $Q_{i,\pm} \sim \chi^2((r-1)(c-1))$, where χ^2 is the χ^2 distribution and r, c the degrees of freedom. These degrees of freedom refer to the number of categories for the variables. In the study design of this Thesis, each tail is compared to the inlier population separately. This implies that $r = 2$ is the number of categories on the metabolite distribution (positive tail, inlier; negative tail, inlier). Each gene is tested independently; each gene contains two two-categories: carrier and non-carrier. Thus for each candidate gene i, we have the following quantities,

$$
\begin{aligned}
E_i[+] &= n_+ f_i, \\
E_i[-] &= n_- f_i, \\
Q_{i,+} &= (O_{i,+} - E_i[+])^2 / E_i[+], \\
Q_{i,-} &= (O_{i,-} - E_i[-])^2 / E_i[-].
\end{aligned}
$$

Above, f_i is the frequency of variant carriers within the inlier population, n_+, n_- are the sizes of the tail populations and $Q_{i,+}, Q_{i,-}$ are the test quantities measuring association.

7.3 Metabolic variation and outliers

The state of metabolism varies between individuals as well as within individuals over time. Numerous factors may explain this variation: genetic, environmental and developmental factors have been investigated [37, 38]. Typically, extreme observations or outliers have been discarded or ignored in these studies. However, it is known that IEM result in extreme metabolite levels in tissue or biofluid. This shows that genetic variation might explain some of these discarded outliers. The robustness of these outliers observations presents a challenge. By definition, outliers are observed sparsely, so biological or measurement noise may explain their occurence.

There are several statistical definitions of outliers [39]. In this Thesis, outliers are defined as observations in the tail of the univariate Gaussian distribution. Given standardized metabolite concentrations z, outliers $z_{outlier}$ are defined as:

$$z \sim N(0,1) \tag{1}$$

$$|z_{outlier}| \geq z_c \tag{2}$$

$z_c > 0$ is the threshold of extremity.

This definition can be applied sensibly if the distributions examined are unimodal and symmetric. The larger the deviation from these conditions, the less motivated this definition is. Figure 13 shows an example of this.

Figure 13: Outlier threshold marked on three distributions. The sensibility of this definition varies: A. The distribution, which is symmetric and unimodal. Outliers are sensible and well-defined. B. The distribution is unsymmetric. The tails are unequal in size, thus a common threshold is questionable. C. The distribution is bimodal and skewed. The outliers are ill-defined.

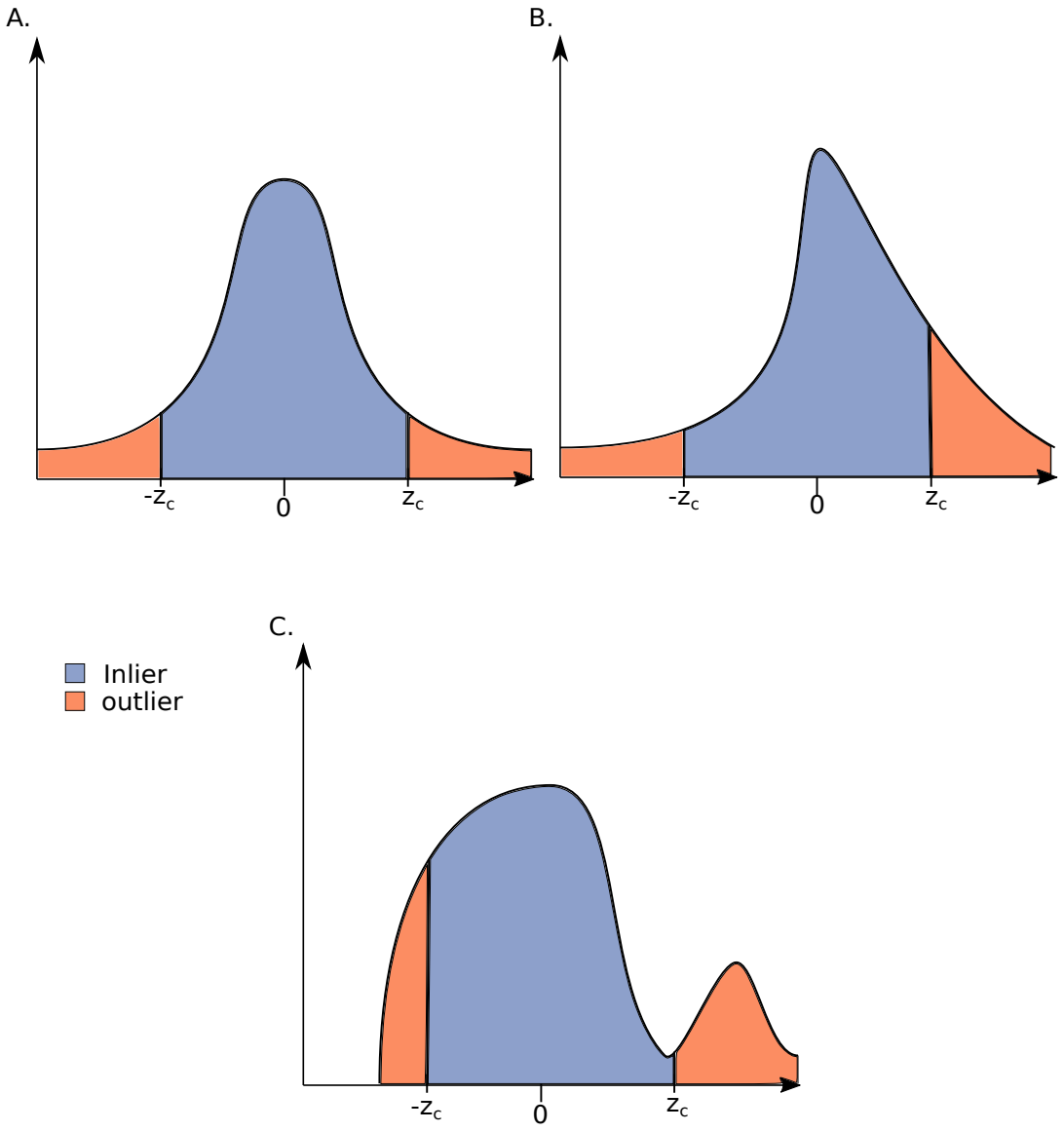
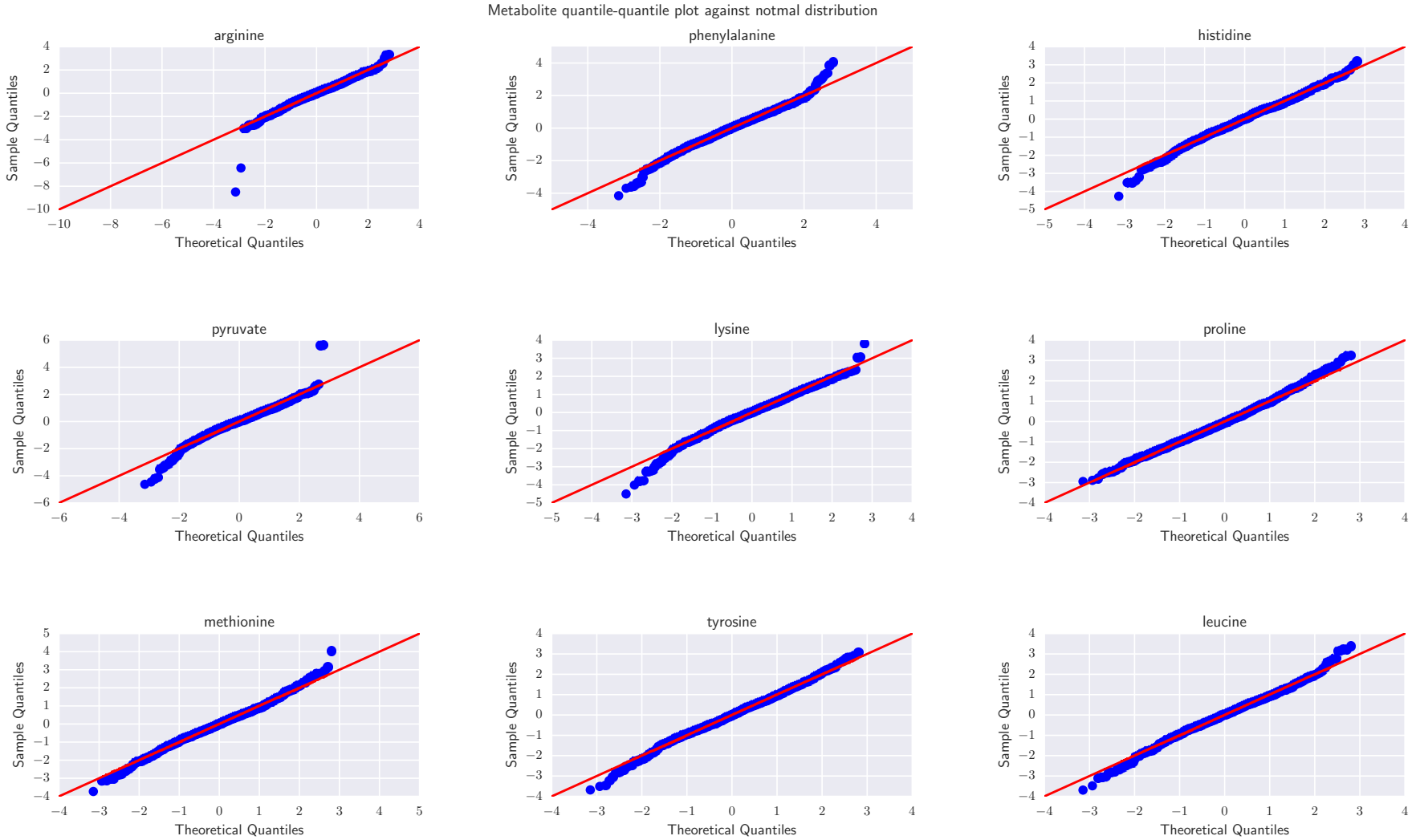


Figure 14: QQ-plot of IEM metabolites against the standardized normal distribution



QQ-plots can be used to characterize the properties of distributions in parameter-free manner. The QQ-plot compares the number of observations per quantile between a reference distribution and an empirical distribution. In Figure 14, the IEM metabolite distributions have been plotted against the standard normal distribution. Notably, all metabolite distributions conform to the reference distribution within the central quantiles $z \in [-2, 2]$. However, the tails are markably skewed and each distribution exhibits different rates of deviation from the reference distribution. This implies that optimally z_c would be chosen individually per distribution per tail. However, this might introduce comparability issues between metabolites. Based on these observations and Figure 14, a reasonable range for a common z_c would lie in the interval $[2, 3]$.

The choice of z_c presents a potential trade-off; lower z_c provides higher sample size, but may in principle lower the power of the employed test [40]. Testing the effect of z_c on power is method dependent and therefore out of the scope of this Thesis. Therefore $z_c = 2$ remains as the natural choice.

7.4 Cohorts

The study subjects were sampled from two cohorts: COROGENE and Predict-CVD cohorts. The COROGENE-study cohort consists of case and control subjects from the acute coronary syndrome [11]; however only control subjects were selected for the study of this Thesis. This Thesis. These subjects participated in the FINRISK 07 survey [41] The subjects The Predict-CVD cohort consists of case and control subjects from a prospective cardiovascular disease (CVD) study of Vartiainen *et al.* [41]. All participated in the FINRISK survey of 97.

Table 1: Summary statistics of the subjects from the COROGENE cohort

-	male	female	total
N	278	236	514
mean BMI($\frac{kg}{m^2}$)	27.0 ± 3.8	26.7 ± 5.4	26.8 ± 4.7
mean age(a)	52.6 ± 13.6	50.8 ± 13.9	51.6 ± 13.8

Table 2: Summary statistics of the subjects from the PredictCVD cohort

-	male	female	total
N	404	231	635
mean BMI($\frac{kg}{m^2}$)	27.5 ± 3.9	26.3 ± 4.4	27.0 ± 4.1
mean age(a)	50.7 ± 13.8	51.0 ± 13.3	50.9 ± 13.6

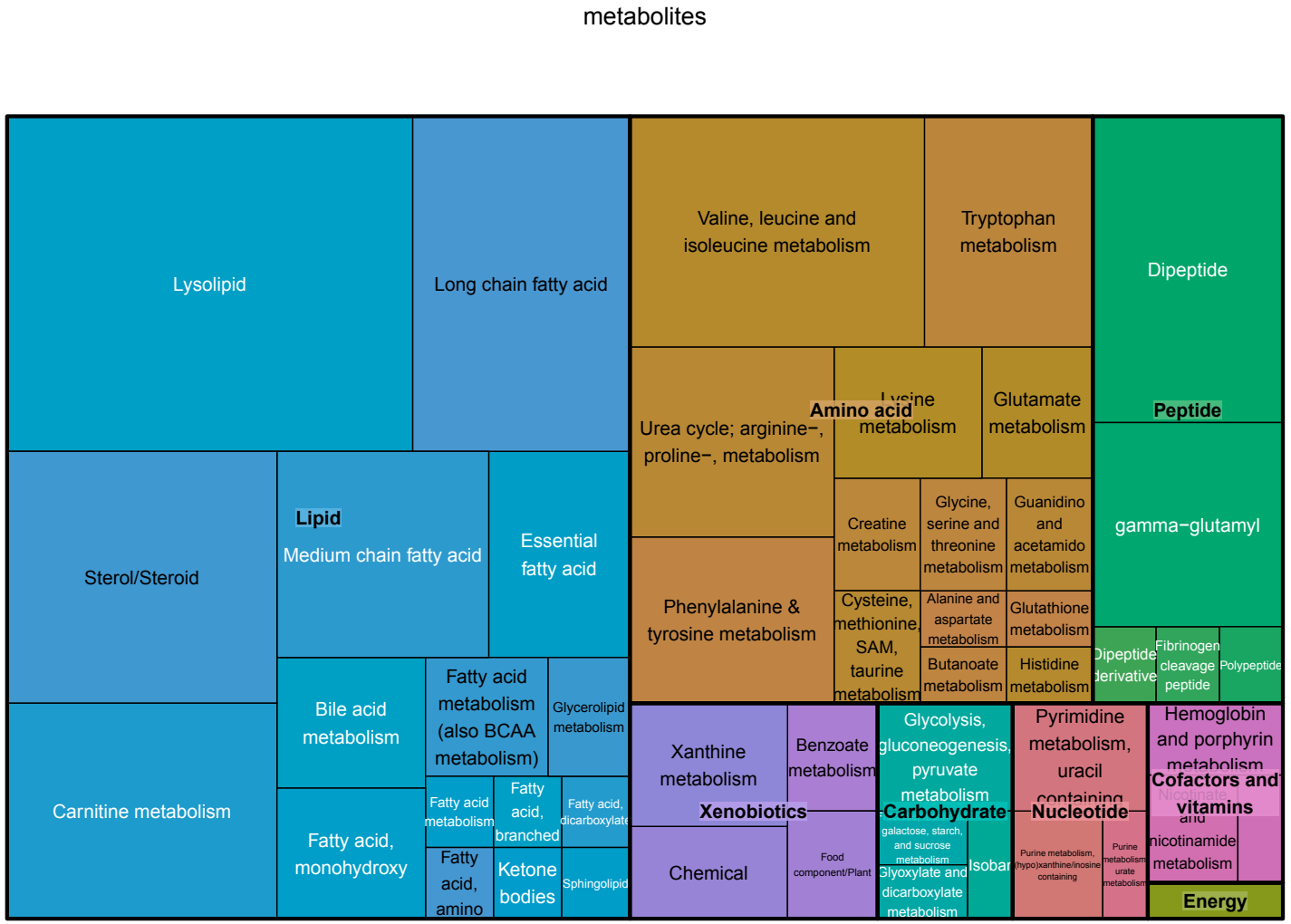
7.5 Metabolite measurement

The subjects were asked to fast for four hours prior to blood sampling. The blood was extracted from a vein in the antecubital fossa. Then, the sampled blood was kept for 30 min at room temperature and centrifuged for 11 min. at 2200g using a

Hettich Rotofix 32 centrifuge. Finally, the plasma and the serum were aliquoted and frozen either with dry ice or to -20 deg C using a freezer. Whole blood samples used for DNA extraction were freezed in the latter manner.

Sample preparation was conducted as described Evans *et al.* [42] by Metabolon, Inc.

Figure 15: Metabolites with more than 50% and known name divided into eight superpathways and 62 subpathways. The area enclosed by each box in proportional to the fraction of metabolites in the pathway.



In total, 647 metabolites were reliably identified with 281 previously unknown. These metabolites can be classified into eight categories: amino acids, carbohydrates, cofactors & vitamins, energy metabolites, lipids, nucleotides, peptides and xenobiotics. These superpathways are further subdivided to 62 subpathways.

Metabolites were selected based on number of missing values. Metabolites with more than 50% missing values were discarded. Figure15 summarizes the set of metabolites selected in preprocessing.

7.6 Genotype measurement

The COROGENE cohort was genotyped using Illumina 670k array and the PredictCVD using Illumina 710k. Imputation;

Due to the complex ramifications of sex chromosome inheritance, only variants in autosomal chromosomes shall be studied in this Thesis. Chosen variants fulfilled three conditions : possessing $MAF \leq 5$ according to Exome aggregation consortium (EXAC), lying in the coding region of a gene of interest and predicted by VEP (see next Subsection) to impact the resulting protein moderately or highly. In total X variants fulfilled these conditions. See appendix for additional information. Genotype data was processed using plink (v.1.90) and QCTOOL.Individuals with haplotype confidence for a variant less than 90% were declared missing with respect to said variant.

Genes of interest consist of two sets: those associated with IEM, and those implied by existing mGWAS-hits from Shin.et al. [34].

The KEGG-database [4] provided the IEM gene-metabolite pairs.

7.7 Statistical analysis and annotation resources

The bulk of statistical computation was performed in python 3.5.2 using Pandas package (version 0.18.1) and Scipy (version 0.17.1) on Mac (version OSX 10.5.6).

Metabolite levels were log transformed and thereafter standardized. The transformed measurements were then corrected for age, sex, BMI, cohort membership and fast length via linear regression using the package scikit-learn (v.0.17.1). VEP was used to predict variant effect. KEGG (Release 89.0+/03-22) [4] was used to choose IEM metabolites as well as the associated genes. The database was accessed using the Bioservices (v. 1.5.2) python package. All genomic distance were computed using pybedtools (v. 0.7.5) and Ensembl (GRCh37.74).

8 Results

The study design employed by this Thesis subdivides into two phases: Preprocessing and feature selection, and enrichment testing. Summarily, the first stage consists of selecting metabolites and candidate genes of interest. Next, distributions of selected metabolites are subdivided into inlier-outlier groups and candidate genes are

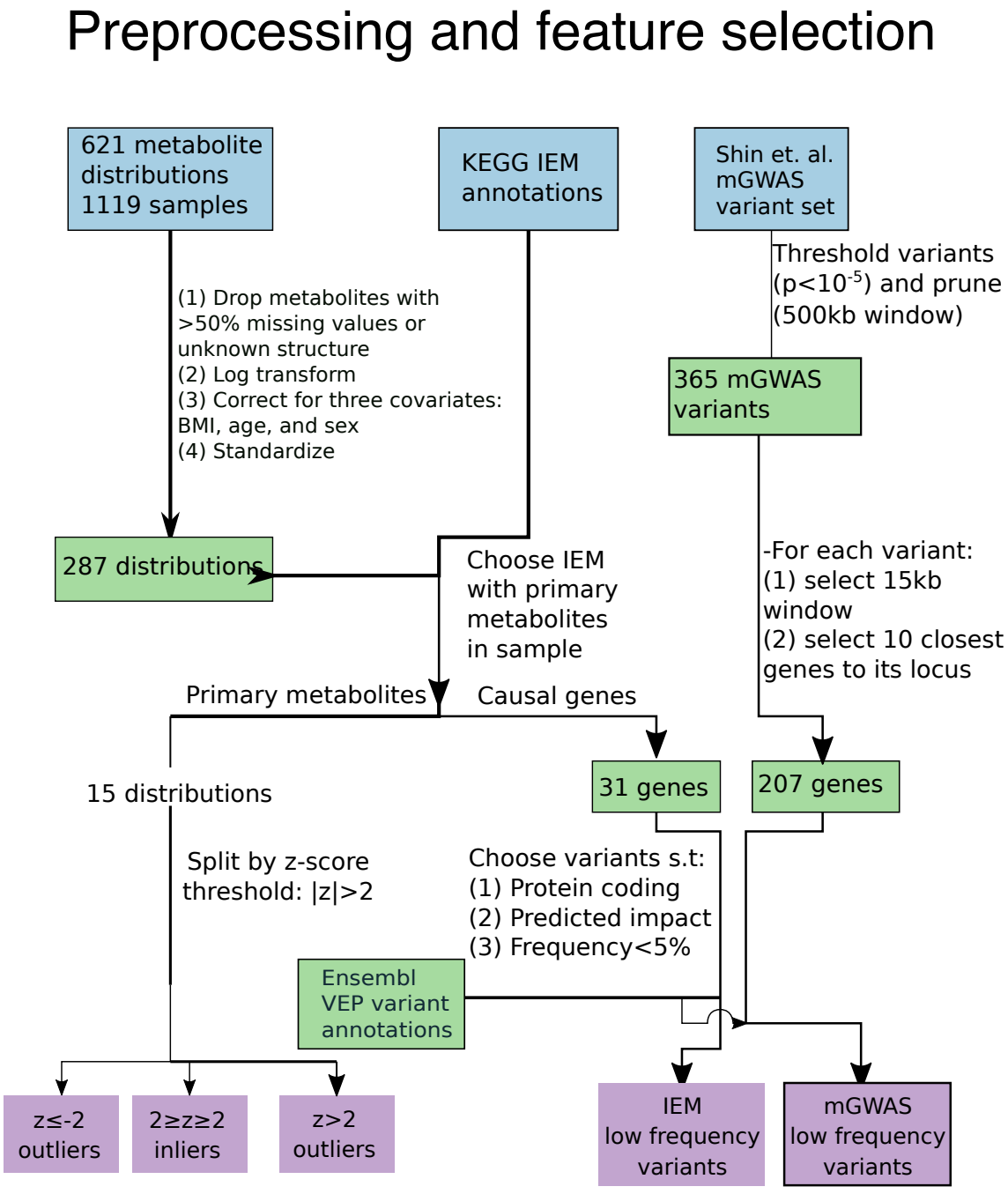


Figure 16: The detailed flow of the study design.

examined for low-frequency variants of interest. The detailed flow of data selection and preprocessing are outlined in Figure 8.

The enrichment testing in turn consists of comparing combined frequencies of variants per gene-region between outlier and inlier groups. Outlier groups consist of the positive and negative tail of the distribution and the inlier group the remaining center. The enrichment was tested using the standard χ^2 test comparing total variant frequency per gene of the inlier against both outlier groups.

8.1 Candidate genes

The IEM of interest were chosen based on two condition: first, they must be linked in KEGG to metabolites within the sample and second, their causative gene must lie within a autosomal chromosome.

As previously explained, we chose metabolites exhibiting extreme variation ($|z| > 2$). The number of these metabolites is 9. However, one metabolite (serotonin (5HT)) was associated with a gene in the X chromosome and thus was excluded from analysis.

Table 3: KEGG IEM with metabolites present in sample

	name	KEGG ID	linked metabolite	Type	OMIM
H00165	Tyrosinemia	H00165	tyrosine	Organic acidaemia	276700 276600 276710 140350
H00167	Phenylketonuria;	H00167	phenylalanine	Organic acidaemia	261600 261630 233910 261640 264070
H00171	Histidinemia	H00171	histidine	Organic acidaemia	235800
H00172	Maple syrup urine disease	H00172	leucine	Organic acidaemia	248600 246900 615135
H00184	Hypermethioninemia	H00184	methionine	Organic acidaemia	250850 180960 606664 614300
H00186	Hyperargininemia	H00186	arginine	Organic acidaemia	207800
H00188	Hyperlysinemia	H00188	lysine	Organic acidaemia	238700
H00190	Hyperprolinemia	H00190	proline	Organic acidaemia	239500 239510
H00469	Mitochondrial DNA depletion syndrome	H00469	pyruvate	Mitochondrial disease	603041 609560 251880 203700 613662 612073 256810 271245 612075 245400 221350 615084 615418 615471 617156

Table 4: IEM associated autosomal genes

	Gene names
Histidinemia	HAL
Hyperargininemia	ARG1
Hyperlysinemia	AASS
Hypermethioninemia	MAT1A,AHCY,GNMT,ADK
Hyperprolinemia	PRODH2,ALDH4A1
Maple syrup urine disease	BCKDHA,BCKDHB,DBT,DLD,PPM1K
Mitochondrial DNA depletion syndrome	TYMP,TK2,DGUOK,MDP1,SUCLA2 ,MPV17,C10orf2,RRM2B,SUCLG1 ,AGK,MGME1,SLC25A4,FBXL4,TFAM
Phenylketonuria;	PAH,QDPR,GCH1,PTS,PCBD1
Tyrosinemia	FAH,TAT,HPD

Table 5: The summary of IEM candidate genes: each is metabolite paired with corresponding causal IEM genes.

	gene names
histidine	HAL
leucine	BCKDHA,BCKDHB,DBT,DLD,PPM1K
phenylalanine	PAH,QDPR,GCH1,PTS,PCBD1
tyrosine	FAH,TAT,HPD
lysine	AASS
methionine	MAT1A,AHCY,GNMT,ADK
arginine	ARG1
proline	PRODH,ALDH4A1
pyruvate	TYMP,TK2,DGUOK,MDP1,SUCLA2,MPV17,C10orf2,RRM2B...

Tables 3 and 4 diplay the chosen IEM and their causal genes. The metabolite are finally paired with corresponding IEM genes in 5

metabolite	SNP	distance (bp)	gene name
arginine	rs511304	2555	NAALADL2
	rs6479442	5706	FGD3
	rs6059244	9842	DEFB121
	rs2781668	495,1843	ARG1,MED23
	rs2293683	0,9194,10107,14263	FARSA,SYCE2,CALR,GCDH
	rs12985777	430,1429,6575	JSRP1,OAZ1,C19orf35
	rs12714263	2493	CLIP4
	rs12061159	5173	RGS21
	rs7036499	8998,14259	SLC25A25,NAIF1
	rs837763	2110	PIEZO1
histidine	rs715	0	CPS1
	rs7014133	1085	ADCY8
	rs1527683	3124,3124	FPGT-TNNI3K,TNNI3K
	rs3733588	0	SLC2A9
lysine	rs10165613	1706	SCN9A
	rs10491431	120	UGT3A1
	rs12268257	2865	GPR158
	rs12927959	7364	GFOD2
	rs1364344	12483	SMPD3
	rs1402325	427,4391,5449,5634,8548	ZNF576,SRRM5,IRGQ,ZNF428,L34079.2
	rs1714369	5711	INSC
	rs3024547	781	IL4R
	rs6587731	2379,10084	PRR9,LELP1
	rs7583413	345	C2orf83
	rs1179979	2765	SLC25A21
methionine	rs11925382	83	DNAJC13
	rs17478241	1474	ARHGAP5
	rs3740996	0,9637	TRIM5,TRIM22
	rs4693555	1823	COPS4
	rs6008770	0,8420	TRMU,CELSR1
	rs6533609	4719	ALPK1
	rs2066415	14534	KIAA2026
phenylalanine	rs1565225	218	RALYL
	rs2953014	1013	NF1
	rs937475	191	PAH
	rs10050054	2604	RP11-389E17.1
proline	rs893175	8009	CBLN1
	rs7187836	10976	C16orf47
	rs715544	1408,2483,13471	DGCR14,GSC2,TSSK2
	rs6798346	12759	PPP4R2
	rs2935623	12048,14587	C5orf38,IRX2
	rs175866	1264	MTHFD1L
	rs10512634	1000	UGT3A1
	rs10197940	12478	RIF1
	rs7560860	5588	EML4
pyruvate	rs1133034	1192	ATP11B
	rs11259923	3849	ADAMTSL3
	rs10770353	5090,11184	PIK3C2G,RERGL
	rs7751620	14517	MCHR2
	rs4648464	11228	PRDM16
tyrosine	rs10033622	2605	SORBS2
	rs936632	3108	FAM159A
	rs906466	11416	DLGAP4
	rs8057124	9496	IST1
	rs7834337	10032	ERICH1
	rs7151615	2055	TTC5
	rs4808739	10288	IL12RB1
	rs12891	0	CERS4
	rs11243929	1167,8576	TSC1,C9orf9
	rs10516469	6642,7683	PPP3CA,AP001816.1
	rs10500559	375	PHLPP2
	rs11625663	299,2508	SNW1,SLIRP

Table 6: The selected mGWAS-variants which provided candidate genes and those genes. For brevity, the variants and candidate genes of the metabolite leucine are listed in Appendix A.

The mGWAS-candidate genes were selected based on Shin *et al.* [3] results. The published SNP-list was thresholded with $p \leq 10^{-5}$ and LD-pruned over a 500kb window as described in 7.1.

Finally, each selected variant produced candidate genes listed in **??**. Genes were selected from within a 15kb window centered around each chosen variant. At most ten genes were chosen per window. The pruning and threshold parameters were chosen to produce a comparable number of candidate genes to the IEM candidate genes. The selected mGWAS variants and the associated genes are listed in 8.1.

8.2 Selected variants

Chosen variants for both candidate gene sets were chosen based on two criteria,

- If the any minor alleles satisfy $MAF \leq 5\%$ in according to ExAc estimated frequency in finnish population.
- If the variant lies within the coding region (i.e, VEP annotated "BIOTYPE" is "protein_coding")

Table 7: The number of low-frequency variants selected per IEM candidate gene.

metabolite	gene	num.variants
arginine	ARG1	3
	Total	3
phenylalanine	GCH1	3
	PAH	11
	PCBD1	1
	PTS	5
	QDPR	6
	Total	26
pyruvate	AGK	3
	C10orf2	14
	DGUOK	18
	FBXL4	14
	MDP1	5
	MGME1	8
	MPV17	12
	OPA1	22
	RRM2B	5
	SLC25A4	4
	SUCLA2	4
	SUCLG1	21
	TFAM	12
	TK2	11
	TYMP	11
	Total	164
proline	ALDH4A1	20
	PRODH	25
	Total	45
histidine	HAL	16
	Total	16
methionine	ADK	18
	AHCY	10
	GNMT	6
	MAT1A	12
	Total	46
leucine	BCKDHA	10
	BCKDHB	9
	DBT	8
	DLD	6
	PPM1K	3
	Total	36
lysine	AASS	15
	Total	15
tyrosine	FAH	10
	HPD	8
	TAT	5
	Total	23

Table 8: The number of low-frequency variants selected per mGWAS candidate gene. The number of variants chosen for leucine candidate genes are listed in Appendix A.

metabolite	gene	num.variants
arginine	ARG1	6
	C19orf35	49
	CALR	28
	CLIP4	16
	DEFB121	1
	FARSA	70
	FGD3	51
	GCDH	35
	JSRP1	28
	MED23	28
	NAALADL2	16
	NAIF1	6
	OAZ1	49
	PIEZO1	106
	RGS21	3
	SLC25A25	45
	SYCE2	56
	Total	593
phenylalanine	NF1	28
	PAH	11
	RALYL	7
	Total	46
pyruvate	ADAMTSL3	23
	ATP11B	12
	MCHR2	4
	PIK3C2G	60
	PRDM16	39
	RERGL	4
	Total	142
proline	C16orf47	12
	C5orf38	15
	CBLN1	2
	DGCR14	63
	EML4	32
	IRX2	18
	MTHFD1L	14
	PPP4R2	7
	RIF1	54
	TSSK2	9
	UGT3A1	30
	Total	256
histidine	ADCY8	18
	CPS1	33
	FPGT-TNNI3K	58
	SLC2A9	12
	Total	121
methionine	ALPK1	48
	ARHGAP5	9

The number of variants varied by metabolite and gene. Table 7 summarizes this information for IEM and Table 8for mGWAS.

8.3 Enrichment analysis

The complete outline of the performed test is provided by RVAS section 7.2. Briefly, the enrichment test examines the equality of the proportion of variant carrier per tail group (p_+, p_-) against the proportion in the inlier population (p_{inlier}). More concisely,

Null-hypothesis(1) $p_+ = p_{inlier}$,

Null-hypothesis(2) $p_- = p_{inlier}$,

Alternative hypothesis: $p_+ \neq p_{inlier}$ or $p_- \neq p_{inlier}$.

Both tails are tested independently. Within the two comparison, each gene constitutes an indepent test. The number of test varies between metabolites and hence the multiple testing corrected p-value threshold varies as well. The correction here used here is the Bonferroni correction.

metabolite	gene counts	Bonferroni p-value
phenylalanine	4	0.012500
arginine	17	0.002941
proline	12	0.004167
tyrosine	15	0.003333
lysine	16	0.003125
methionine	9	0.005556
pyruvate	6	0.008333
histidine	5	0.010000
leucine	114	0.000439

Table 9: The number of IEM-genes per metabolite. The number of genes corresponds to the number of χ^2 tests performed per metabolite. To account for multiple tests, the nominal p-value $p = 0.05$ is divided by the number of tests.

metabolite	gene counts	Bonferroni p-value
phenylalanine	4	0.012500
arginine	17	0.002941
proline	12	0.004167
tyrosine	15	0.003333
lysine	16	0.003125
methionine	9	0.005556
pyruvate	6	0.008333
histidine	5	0.010000
leucine	114	0.000439

Table 10: The number of mGWAS-genes per metabolite. The number of genes corresponds to the number of χ^2 tests performed per metabolite. To account for multiple tests, the nominal p-value $p = 0.05$ is divided by the number of tests.

The results of the association test are displayed in abbreviated form in this section. For the positive tail $z > 2$, the results are contained in Table 13 for the IEM variants and mGWAS variants 14. The respective negative tail results are contained in Table 11 and Table 14.

The Tables 8.3 and 8.3 list the Bonferroni corrected p-value thresholds per metabolite. The IEM-results are not significant at the nominal threshold $p \leq 0.05$. The mGWAS-results contain nominally significant results, but none are significant after Bonferroni correction. In the wider context, the results do not support the exact hypothesis, that coding variant are not enriched within the tail subpopulations. This in turn does not support the qualitative hypothesis regarding the intermediate

Table 11: Top ten associations of IEM candidate gene variants from negative tail($z < -2$). There are no nominally significant ($p \leq 0.05$) tests. The number of observed carriers in the negative tail is indicated by n_- and the number of expected carriers is indicated by $E[n_-]$. The p-value of the χ^2 test for enrichment is indicated by $p(-)$

	n_-	$E[n_-]$	n_{inlier}	Q(-)	p(-)	metabolite	num. variants
PRODH	5.0	5.501799	322.0	0.045767	0.169400	proline	25
QDPR	1.0	0.801259	27.0	0.049295	0.175705	phenylalanine	6
MAT1A	1.0	1.262590	52.0	0.054613	0.184777	methionine	6
HPD	3.0	3.525180	112.0	0.078241	0.220305	tyrosine	8
PCBD1	0.0	0.089029	3.0	0.089029	0.234584	phenylalanine	1
AGK	0.0	0.107914	4.0	0.107914	0.257468	pyruvate	3
FBXL4	3.0	2.428058	90.0	0.134724	0.286417	pyruvate	14
PAH	1.0	1.454137	49.0	0.141830	0.293531	phenylalanine	11
PTS	2.0	1.513489	51.0	0.156389	0.307497	phenylalanine	5
SLC25A4	5.0	4.127698	153.0	0.184343	0.332333	pyruvate	4

Table 12: Top ten associations of mGWAS candidate gene variants from negative tail ($z < -2$). The number of observed carriers in the negative tail is indicated by n_- and the number of expected carriers is indicated by $E[n_-]$. The p-value of the χ^2 test for enrichment is indicated by $p(-)$

	n_-	$E[n_-]$	n_{inlier}	Q(-)	p(-)	metabolite	num. variants
SRRM4	8.0	7.884892	274.0	0.001680	0.032698	leucine	14
CELF4	4.0	3.856115	134.0	0.005369	0.058411	leucine	10
IRGQ	1.0	1.079137	40.0	0.005803	0.060724	lysine	3
NRG3	3.0	2.848921	99.0	0.008012	0.071322	leucine	10
SLC25A25	2.0	2.135791	95.0	0.008633	0.074030	arginine	15
CERS4	1.0	1.101619	35.0	0.009374	0.077129	tyrosine	10
TRNT1	3.0	2.820144	98.0	0.011470	0.085290	leucine	10
FGD3	6.0	6.272482	279.0	0.011837	0.086637	arginine	17
IRF2	4.0	3.769784	131.0	0.014059	0.094384	leucine	11
COPS4	5.0	5.317446	219.0	0.018951	0.109494	methionine	6

Table 13: Top ten associations of IEM candidate gene variants from positive tail($z > 2$). There are no nominally significant ($p \leq 0.05$) tests. The number of observed carriers in the positive tail is indicated by n_+ and the number of expected carriers is indicated by $E[n_+]$. The p-value of the χ^2 test for enrichment is indicated by $p(+)$

	n_+	$E[n_+]$	n_{inlier}	Q(+)	p(+)	metabolite	num. variants
HAL	2.0	2.179856	101.0	0.014840	0.096957	histidine	16
ADK	1.0	0.836331	31.0	0.032030	0.142038	methionine	8
PTS	1.0	0.825540	51.0	0.036869	0.152267	phenylalanine	5
PAH	1.0	0.793165	49.0	0.053936	0.183650	phenylalanine	11
DLD	1.0	1.282374	62.0	0.062178	0.196914	leucine	6
AGK	0.0	0.071942	4.0	0.071942	0.211471	pyruvate	3
DGUOK	2.0	2.428058	135.0	0.075465	0.216460	pyruvate	6
AASS	1.0	1.456835	90.0	0.143254	0.294933	lysine	15
ALDH4A1	13.0	14.511691	489.0	0.157474	0.308507	proline	19
BCKDHA	2.0	1.509892	73.0	0.159088	0.310002	leucine	10

Table 14: Top ten associations of mGWAS candidate gene variants from positive tail($z > 2$). The number of observed carriers in the positive tail is indicated by n_+ and the number of expected carriers is indicated by $E[n_+]$. The p-value of the χ^2 test for enrichment is indicated by $p(+)$

	n_+	$E[n_+]$	n_{inlier}	Q(+)	p(+)	metabolite	num. variants
PHACTR2	4.0	4.012590	194.0	0.000040	0.005015	leucine	12
SDHD	1.0	1.013489	49.0	0.000180	0.010691	leucine	5
ADAM12	4.0	4.053957	196.0	0.000718	0.021379	leucine	20
TSC22D2	2.0	1.902878	92.0	0.004957	0.056130	leucine	11
RIN3	1.0	1.075540	52.0	0.005305	0.058065	leucine	10
CERS4	1.0	0.912770	35.0	0.008336	0.072748	tyrosine	10
FAM149A	2.0	2.151079	104.0	0.010611	0.082044	leucine	13
VAX2	1.0	1.116906	54.0	0.012237	0.088082	leucine	5
MAP2K6	1.0	1.116906	54.0	0.012237	0.088082	leucine	7
SNW1	1.0	1.121403	43.0	0.013143	0.091272	tyrosine	10

impact of low-frequency genetic variation stated earlier in the Section 1.

9 Discussion

The results do not support to the hypothesis of intermediate impact. More concisely, the hypothesis whereby low-frequency genetic variation constitutes a novel regime of impact on metabolic traits besides the known regimes of common and Mendelian variation. This outcome is possible through two scenarios: either the impact remains undetected in this Thesis or low-frequency does not constitute an independent regime by impact.

Both scenarios can be analyzed via four fields illustrated in Figure17. The division of the Thesis' study into metabolic and genomic aspects is self-evident; however the division into sampling and feature selection aspects warrants further elaboration. Features refer to the set of relevant variables. Selecting candidate genes reduces the number of variables comprising the genotype. Analogously, selection and transformation of metabolite distributions define the relevant metabolic phenotypes.

9.1 Genotype sampling aspect— study design and association testing

The most direct factor influencing detection is the choice of statistical test. The intrinsic properties of the χ^2 test either boost or dampen its power in a given experimental design. Most likely, the defining property of the χ^2-test is its non-robustness to low number of expected observations in any group. The heuristic threshold of minimum group size is 5 samples [36]. Simple modifications to test this assumption would be to select candidate genes with more plausible LOF low-frequency variants leading to higher carrier count on expectation. However, it is unclear whether this leads to bias and to which direction.

The usage of the test in the Thesis' design presents an additional confounding factor. Namely, that it assumes coding variants have a similar impact in direction and magnitude. This could be potentially ameliorated by modifying the study design by stratifying the variants by expected impact size or direction. The expected impact

Figure 17: A fourfold division of the study of this Thesis. On the level of theory, the Thesis involves two domains: metabolic and genomic. On the level of study design or experiment, the two domains are feature and sampling. The feature domain dictates what is being observed. The sampling domain dictates what are observations. Their intersections comprise four distinct aspects of the Thesis study.

	Genetic	**Metabolic**
Feature	Candidate gene	Metabolite distribution
Sample	genotype	outlier

could be estimated from mutation type (missense mutation,non-sense mutation) or minor allele frequency. Then each strata would constitute an independent χ^2 test.

An other option would involve choosing an alternative test. Lee *et al.* report a number of rare-variant tests, which account for some of these heterogeneities [35]. Some of these test may simultaneously correct for the impact of common variation as well [44] and can be adapted through weighting-schemes.

9.2 Genetic feature aspect— gene selection

The choice of IEM candidate genes was largely dictated by available annotation resources. The KEGG [4] database provided both IEM genes and metabolites. Grouped according to the criterion presented by Salmi *et al.* [31] (see also section 6), the IEMs provided by KEGG are all organic acidemias. This might incur a selection bias.

The choice of mGWAS candidate genes provides for degrees of freedom. The choice of single nucleotide polymorphism threshold p-values, pruning distances, gene-window length and maximum number of genes chosen from the window affect the number and chromosomal distribution of candidate genes significantly. The current choice of parameters normalizes the number of mGWAS candidate genes to match the number of IEM genes. Curiously, the the two candidate gene sets have only one common gene: PAH.

This observation raises a crucial question: to what extent is it a consequence of the sets' origins (mGWAS, IEM)? The above mentioned parameter-choice issue and the statistical nature of mGWAS-results might inject noise to the process.

Brodie *et al.* [33] have investigated the effect of gene window length when associating mGWAS-variants to causative genes. Using a method presented in the article, they discovered that not only did optimal length vary by phenotype, but that for some single nucleotide polymorphism the length was exceedingly long (500kb compared to this Thesis' 10kb). This method linked each single nucleotide polymorphism to a pathway and discovered a set plausible candidate gene by distance.

This method of generating candidate gene from single nucleotide polymorphisms eliminates the four parameters and provides additional context for the selected genes. Thus, it is more robust without foreseeably reducing the benefits of using mGWAS-variants.

The gene function and structure might affect its susceptibility to LOF. This dimension could be accounted for by studying existing IEM genes and the impact of variation on their chemical activity. Blau *et al.* [45] have studied the causal gene, PAH, of the IEM phenylketonuria.

9.3 Metabolic sampling aspect— outlier definition

Canonically, outliers are interpreted as observations not generated by the process of interest. In a carefully constructed, controlled experiment, typically only one process is of interest; the rest comprise measurement noise. Thus samples conforming to expected tendencies of this process can be justifiably selected and non-conforming

outlier samples discarded. However, when outliers are generated by the process of interest, this distinction is inapplicable. This raises the question: how to distinguish extreme observations attributable to relevant processes from those attributable to noise? Concretely this issue manifests itself in this Thesis in the choice of outlier thresholds. While a high threshold dampens measurement noise, it reduces sample size.

The assumed type of the distributions impacts the choice critically. An elementary example is attributing observations from a fat-tailed distribution to a thin-tailed distribution. More concretely, if the observations are generated by a Student-t distribution, then assuming a normal distribution leads to a significant number of outliers. Since the Thesis dataset contains a range of skewed and kurtotic distributions which persist even after log-transformation, this issue is relevant to the the results. Furthermore, the literature on expected distribution type and shape is lacking.

One notable phenomenon regarding extreme observations in high-dimensional data, is the curse of dimensionality. The curse of dimensionality refers to multiple linked phenomena; however, the phenomenon of data-snooping bias applies acutely [46]. The data-snooping bias states that given any individual sample, the probability of this individual being an outlier increases as the number of dimensions increases. As the number of tested metabolite distributions increases, the probability of sampling an individual who is not an outlier with respect any metabolite drastically drops. The bias affects unsupervised (model-free) outlier detection. Thus in principle, the bias could be mitigated or at least accounted for by explicitly assuming a model.

Thus, a statistical framework to quantify the number of expected outlier and their magnitude is needed. Such a framework could rely on domain agnostic methods such as extreme-value analysis [47] or biological models such as the Gaussian graphical machines studied by Krumsiek *et al.* [48].

Additionally, other mathematical formulations of outliers than the distribution based might prove more robust to these phenomena [39], though extensive work must done to ascertain this.

Furthermore, the measurement process of metabolite levels might affect distributions as well. Differences of detection thresholds might introduce spurious outliers. Specifically, this applies to metabolites with significant numbers of missing observations, since missingness can be attributed to concentrations below detection threshold.

As mentioned before, many of the metabolite distribution are kurtic. Since kurtosis refers to the tendency of the distribution to produce outliers [49], analytic sources of it undermines the biological relevance of outliers. This could be investigated through studying whether mass-spectrometric or chromatographic attributes of compounds predict kurtosis significantly.

9.4 Metabolic sampling aspect-– biological validation

The metabolic state of an individual varies across time. The scale of this variation is subject to the physiological and clinical state of the individual. For example, individuals diagnosed with organic acidemias display serum amino acid contents

orders of magnitude higher from the healthy population.

In the context of this Thesis, outliers are extreme observations with respect to the sampled population. Thus, in principle they might lie well within range of clinical or biological reference levels.

The biochemical validation could be done through comparisons with in silico models. Nijhout *et al.* [50] have shown in silico that the chemical reaction networks of at least certain metabolic pathways are capable of buffering against perturbation innately. They due so partly by virtue of their network structure and by active regulation of reaction fluxes. In the context of the Thesis, their key insight is that this buffering effect mitigates perturbations regardless of origin; that is, external or internal. Thus a loss-of-function in an enzyme or transporter protein might not result in extreme perturbation of individual metabolites, but moderate perturbation of entire metabolic processes. This might thus manifest as a lack of univariate-Gaussian outliers despite biological relevance. This motivates the investigation of the joint-distribution of metabolites — of multidimensional outliers [39].

Contrarily, given the non-essential nature of some metabolites, extreme variation in these metabolites might prove irrelevant.

Thus, the context provided by these models could be used for selection of relevant metabolites or provide for a framework for combining raw metabolites into composite features. Developing and validating such complex models is beyond the scope of this Thesis or any immediate work. However, such models have been published. Cvitanovic *et al.* [51] have published a review on genome scale dynamic models of liver metabolism.

From a clinical perspective, the model for organic acidoses proposes an empirical alternative. Salmi *et al.* [31] present a decomposition of the metabolic consequences of aminoacidemias (see also **??**). Especially secondary metabolic consequences (clinically significant systemic perturbations such as ketosis) might provided a clinical validation of outlying metabolite concentration.

10 Conclusions

This Thesis has studied the impact of low-frequency coding genetic variation on serum metabolite composition in a Finnish sample of the general population. The genes studied for coding-variation were obtained from prior literature (IEM), as well as empirically from mGWAS results. The candidate metabolites were obtained from IEM-literature. According to the hypothesis presented, the carriers of these variants should be enriched in the tail population of metabolite distributions. Although the results of this Thesis do not confirm the hypothesis, it serves as proof-of-concept for a systemic approach towards the intersection of the clinical, biochemical and statistical life-sciences.

By integrating further clinical insight from IEM, of biochemical pathway knowledge and statistical theory of outliers, the hypothesis presented can be examined from new angles. These ingredients are readily available. However, it is unclear as to which of these directions should occupy the central role in the future. Answering

this question is a matter of interdisciplinary deliberation.

A Leucine genes and variants

Table 15: The selected mGWAS variants and the consequent candidate genes for leucine.

SNP	distance (bp)	gene name
rs37959	4940,6035	RPA3-AS1,GLCCI1
rs3906146	302	LMX1B
rs4073959	615	MTL5
rs4074961	2516	RSPO1
rs3775574	0	IRF2
rs2933284	12464	SLC6A1
rs4681618	4788	TSC22D2
rs440752	5145	REG3G
rs4432749	1379	TBC1D14
rs4593819	0,568	CASP9,DNAJC16
rs2898645	12806	SEPT9
rs4693209	781	PPM1K
rs4758270	1711	TUB
rs4077470	6485	NTF3
rs2454043	2091	ATP6V1C1
rs2722276	528	EPDR1
rs17265788	1391	CES5A
rs1818106	3454,5622	DDI1,PDGFD
rs1822834	12098	ALDH1A2
rs1881909	9906	CPNE4
rs192689	137	NMNAT3
rs1971105	419	EML6
rs211167	4575	FHL5
rs2146498	9482	RIN3
rs2273875	239	TARBP1
rs2275977	388,5000	DERL3,SMARCB1
rs2285137	0,10290	POLDIP3,SERHL2
rs2302423	0	CCDC179
rs4767788	3242	SRRM4
rs2478832	1577	KIAA1462
rs2634945	1201	DDX60L
rs2819318	10229	NOS1AP
rs477992	2793	PHGDH
rs606798	3596	CELF4
rs481781	107	PRKCI
rs7584842	496	GREB1
rs7622687	9577,10424	DNASE1L3,FLNB
rs769630	476	SUMF1
rs7707382	5602	ANKRD33B
rs7912957	2661	ATE1
rs7915	0,1164,6810	ZFYVE19,PPP1R14D,DNAJC17
rs797348	1546	EGLN3
rs7996613	3059	MYO16
rs8131355	4732	SH3BGR
rs849327	12095	JAZF1
rs885561	3949	LIPA
rs891878	8338	MYT1L
rs9403527	14360	PHACTR2
rs9521172	5384	MYO16

Table 16: The number of final selected low-frequency variants per gene for the metabolite leucine

gene	num.variants
ABCC5	266
ADAM12	220
ADORA1	50
ADRA1B	12
AHRR	112
ALDH1A2	45
ANKRD33B	16
ARHGAP21	143
ASB5	80
ASTN2	78
ATE1	110
ATP6V1C1	8
B3GNT3	28
BCMO1	36
BCO2	108
C1orf173	240
C3orf55	14
CASP9	70
CCDC179	12
CELF4	10
CES5A	30
COL4A3	288
CPNE4	70
CRBN	56
CTTNBP2	102
DDI1	24
DDX60L	200
DENND5B	80
DERL3	40
DNAJC16	60
DNAJC17	35
DNASE1L3	140
EGLN3	8
EMCN	48
EML6	396
EPB41L4B	60
EPDR1	24
FAM149A	104
FER1L6	82
FGF12	14
FHL5	15
FLNB	826
FOSL2	27
GBF1	209
GLCCI1	30
GREB1	432
IL18	24
INSL3	12
IRF2	88

References

[1] M. Dawn Teare and Mauro F. Santibañez Koref. Linkage analysis and the study of Mendelian disease in the era of whole exome and genome sequencing. *Briefings in Functional Genomics*, 13(5):378–383, September 2014.

[2] Mark I. Mccarthy, Gonçalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews. Genetics; London*, 9(5):356–69, May 2008.

[3] So-Youn Shin, Eric B. Fauman, Ann-Kristin Petersen, Jan Krumsiek, Rita Santos, Jie Huang, Matthias Arnold, Idil Erte, Vincenzo Forgetta, Tsun-Po Yang, Klaudia Walter, Cristina Menni, Lu Chen, Louella Vasquez, Ana M. Valdes, Craig L. Hyde, Vicky Wang, Daniel Ziemek, Phoebe Roberts, Li Xi, Elin Grundberg, The Multiple Tissue Human Expression Resource (MuTHER) Consortium, Melanie Waldenberger, J. Brent Richards, Robert P. Mohney, Michael V. Milburn, Sally L. John, Jeff Trimmer, Fabian J. Theis, John P. Overington, Karsten Suhre, M. Julia Brosnan, Christian Gieger, Gabi Kastenmüller, Tim D. Spector, and Nicole Soranzo. An atlas of genetic influences on human blood metabolites. *Nature Genetics*, 46(6):543–550, June 2014.

[4] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000.

[5] Edward S. Tobias, Michael Connor, and Malcolm Ferguson-Smith. *Essential Medical Genetics*. John Wiley & Sons, November 2011. Google-Books-ID: 77Dvq_OoMnYC.

[6] Creative Commons — Attribution 4.0 International — CC BY 4.0.

[7] Creative Commons — Attribution-ShareAlike 4.0 International — CC BY-SA 4.0.

[8] Creative Commons — Attribution 2.5 Generic — CC BY 2.5.

[9] Creative Commons — Attribution 3.0 Unported — CC BY 3.0.

[10] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I. Berndt, Michael N. Weedon, Fernando Rivadeneira, Cristen J. Willer, Anne U. Jackson, Sailaja Vedantam, Soumya Raychaudhuri, Teresa Ferreira, Andrew R. Wood, Robert J. Weyant, Ayellet V. Segrè, Elizabeth K. Speliotes, Eleanor Wheeler, Nicole Soranzo, Ju-Hyun Park, Jian Yang, Daniel Gudbjartsson, Nancy L. Heard-Costa, Joshua C. Randall, Lu Qi, Albert Vernon Smith, Reedik Mägi, Tomi Pastinen, Liming Liang, Iris M. Heid, Jian'an Luan, Gudmar Thorleifsson, Thomas W. Winkler, Michael E. Goddard, Ken Sin Lo, Cameron Palmer, Tsegaselassie Workalemahu, Yurii S. Aulchenko, Åsa Johansson, M. Carola Zillikens, Mary F. Feitosa, Tõnu Esko, Toby Johnson, Shamika Ketkar, Peter Kraft, Massimo

Mangino, Inga Prokopenko, Devin Absher, Eva Albrecht, Florian Ernst, Nicole L. Glazer, Caroline Hayward, Jouke-Jan Hottenga, Kevin B. Jacobs, Joshua W. Knowles, Zoltán Kutalik, Keri L. Monda, Ozren Polasek, Michael Preuss, Nigel W. Rayner, Neil R. Robertson, Valgerdur Steinthorsdottir, Jonathan P. Tyrer, Benjamin F. Voight, Fredrik Wiklund, Jianfeng Xu, Jing Hua Zhao, Dale R. Nyholt, Niina Pellikka, Markus Perola, John R. B. Perry, Ida Surakka, Mari-Liis Tammesoo, Elizabeth L. Altmaier, Najaf Amin, Thor Aspelund, Tushar Bhangale, Gabrielle Boucher, Daniel I. Chasman, Constance Chen, Lachlan Coin, Matthew N. Cooper, Anna L. Dixon, Quince Gibson, Elin Grundberg, Ke Hao, M. Juhani Junttila, Lee M. Kaplan, Johannes Kettunen, Inke R. König, Tony Kwan, Robert W. Lawrence, Douglas F. Levinson, Mattias Lorentzon, Barbara McKnight, Andrew P. Morris, Martina Müller, Julius Suh Ngwa, Shaun Purcell, Suzanne Rafelt, Rany M. Salem, Erika Salvi, Serena Sanna, Jianxin Shi, Ulla Sovio, John R. Thompson, Michael C. Turchin, Liesbeth Vandenput, Dominique J. Verlaan, Veronique Vitart, Charles C. White, Andreas Ziegler, Peter Almgren, Anthony J. Balmforth, Harry Campbell, Lorena Citterio, Alessandro De Grandi, Anna Dominiczak, Jubao Duan, Paul Elliott, Roberto Elosua, Johan G. Eriksson, Nelson B. Freimer, Eco J. C. Geus, Nicola Glorioso, Shen Haiqing, Anna-Liisa Hartikainen, Aki S. Havulinna, Andrew A. Hicks, Jennie Hui, Wilmar Igl, Thomas Illig, Antti Jula, Eero Kajantie, Tuomas O. Kilpeläinen, Markku Koiranen, Ivana Kolcic, Seppo Koskinen, Peter Kovacs, Jaana Laitinen, Jianjun Liu, Marja-Liisa Lokki, Ana Marusic, Andrea Maschio, Thomas Meitinger, Antonella Mulas, Guillaume Paré, Alex N. Parker, John F. Peden, Astrid Petersmann, Irene Pichler, Kirsi H. Pietiläinen, Anneli Pouta, Martin Ridderstråle, Jerome I. Rotter, Jennifer G. Sambrook, Alan R. Sanders, Carsten Oliver Schmidt, Juha Sinisalo, Jan H. Smit, Heather M. Stringham, G. Bragi Walters, Elisabeth Widen, Sarah H. Wild, Gonneke Willemsen, Laura Zagato, Lina Zgaga, Paavo Zitting, Helene Alavere, Martin Farrall, Wendy L. McArdle, Mari Nelis, Marjolein J. Peters, Samuli Ripatti, Joyce B. J. van Meurs, Katja K. Aben, Kristin G. Ardlie, Jacques S. Beckmann, John P. Beilby, Richard N. Bergman, Sven Bergmann, Francis S. Collins, Daniele Cusi, Martin den Heijer, Gudny Eiriksdottir, Pablo V. Gejman, Alistair S. Hall, Anders Hamsten, Heikki V. Huikuri, Carlos Iribarren, Mika Kähönen, Jaakko Kaprio, Sekar Kathiresan, Lambertus Kiemeney, Thomas Kocher, Lenore J. Launer, Terho Lehtimäki, Olle Melander, Tom H. Mosley Jr, Arthur W. Musk, Markku S. Nieminen, Christopher J. O'Donnell, Claes Ohlsson, Ben Oostra, Lyle J. Palmer, Olli Raitakari, Paul M. Ridker, John D. Rioux, Aila Rissanen, Carlo Rivolta, Heribert Schunkert, Alan R. Shuldiner, David S. Siscovick, Michael Stumvoll, Anke Tönjes, Jaakko Tuomilehto, Gert-Jan van Ommen, Jorma Viikari, Andrew C. Heath, Nicholas G. Martin, Grant W. Montgomery, Michael A. Province, Manfred Kayser, Alice M. Arnold, Larry D. Atwood, Eric Boerwinkle, Stephen J. Chanock, Panos Deloukas, Christian Gieger, Henrik Grönberg, Per Hall, Andrew T. Hattersley, Christian Hengstenberg, Wolfgang Hoffman, G. Mark Lathrop, Veikko Salomaa, Stefan Schreiber, Manuela Uda, Dawn Waterworth, Alan F. Wright, Themistocles L. Assimes, Inês Barroso,

Albert Hofman, Karen L. Mohlke, Dorret I. Boomsma, Mark J. Caulfield, L. Adrienne Cupples, Jeanette Erdmann, Caroline S. Fox, Vilmundur Gudnason, Ulf Gyllensten, Tamara B. Harris, Richard B. Hayes, Marjo-Riitta Jarvelin, Vincent Mooser, Patricia B. Munroe, Willem H. Ouwehand, Brenda W. Penninx, Peter P. Pramstaller, Thomas Quertermous, Igor Rudan, Nilesh J. Samani, Timothy D. Spector, Henry Völzke, Hugh Watkins, James F. Wilson, Leif C. Groop, Talin Haritunians, Frank B. Hu, Robert C. Kaplan, Andres Metspalu, Kari E. North, David Schlessinger, Nicholas J. Wareham, David J. Hunter, Jeffrey R. O'Connell, David P. Strachan, H.-Erich Wichmann, Ingrid B. Borecki, Cornelia M. van Duijn, Eric E. Schadt, Unnur Thorsteinsdottir, Leena Peltonen, André G. Uitterlinden, Peter M. Visscher, Nilanjan Chatterjee, Ruth J. F. Loos, Michael Boehnke, Mark I. McCarthy, Erik Ingelsson, Cecilia M. Lindgren, Gonçalo R. Abecasis, Kari Stefansson, Timothy M. Frayling, and Joel N. Hirschhorn. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, October 2010.

[11] Satu Vaara, Markku S Nieminen, Marja-Liisa Lokki, Markus Perola, Pirkko J Pussinen, Jaakko Allonen, Olavi Parkkonen, and Juha Sinisalo. Cohort Profile: The Corogene study. *International Journal of Epidemiology*, 41(5):1265–1271, October 2012.

[12] Mark Jobling, Edward Hollox, Matthew Hurles, Toomas Kivisild, and Chris Tyler-Smith. *Human evolutionary genetics*. Garland Science, 2nd edition, 2014.

[13] D Baralle. Splicing in action: assessing disease causing sequence changes. *Journal of Medical Genetics*, 42(10):737–748, October 2005.

[14] Sudha Seshadri. Genome-wide Analysis of Genetic Loci Associated With Alzheimer Disease. *JAMA*, 303(18):1832, May 2010.

[15] William S. Bush and Jason H. Moore. Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12):e1002822, December 2012.

[16] Baninia Habchi, Sandra Alves, Alain Paris, Douglas N. Rutledge, and Estelle Rathahao-Paris. How to really perform high throughput metabolomic analyses efficiently? *TrAC Trends in Analytical Chemistry*, 85:128–139, December 2016.

[17] Darrell D. Marshall and Robert Powers. Beyond the paradigm: Combining mass spectrometry and nuclear magnetic resonance for metabolomics. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 100:1–16, May 2017.

[18] David S. Wishart, Michael J. Lewis, Joshua A. Morrissey, Mitchel D. Flegel, Kevin Jeroncic, Yeping Xiong, Dean Cheng, Roman Eisner, Bijaya Gautam, Dan Tzur, Summit Sawhney, Fiona Bamforth, Russ Greiner, and Liang Li. The human cerebrospinal fluid metabolome. *Journal of Chromatography B*, 871(2):164–173, August 2008.

[19] Christian Gieger, Ludwig Geistlinger, Elisabeth Altmaier, Martin Hrabé de Angelis, Florian Kronenberg, Thomas Meitinger, Hans-Werner Mewes, H.-Erich Wichmann, Klaus M. Weinberger, Jerzy Adamski, Thomas Illig, and Karsten Suhre. Genetics Meets Metabolomics: A Genome-Wide Association Study of Metabolite Profiles in Human Serum. *PLoS Genetics*, 4(11):e1000282, November 2008.

[20] Thomas Illig, Christian Gieger, Guangju Zhai, Werner Römisch-Margl, Rui Wang-Sattler, Cornelia Prehn, Elisabeth Altmaier, Gabi Kastenmüller, Bernet S Kato, Hans-Werner Mewes, Thomas Meitinger, Martin Hrabé de Angelis, Florian Kronenberg, Nicole Soranzo, H-Erich Wichmann, Tim D Spector, Jerzy Adamski, and Karsten Suhre. A genome-wide perspective of genetic variation in human metabolism. *Nature Genetics*, 42(2):137–141, February 2010.

[21] Karsten Suhre, So-Youn Shin, Ann-Kristin Petersen, Robert P. Mohney, David Meredith, Brigitte Wägele, Elisabeth Altmaier, CARDIoGRAM, Panos Deloukas, Jeanette Erdmann, Elin Grundberg, Christopher J. Hammond, Martin Hrabé de Angelis, Gabi Kastenmüller, Anna Köttgen, Florian Kronenberg, Massimo Mangino, Christa Meisinger, Thomas Meitinger, Hans-Werner Mewes, Michael V. Milburn, Cornelia Prehn, Johannes Raffler, Janina S. Ried, Werner Römisch-Margl, Nilesh J. Samani, Kerrin S. Small, H. Erich Wichmann, Guangju Zhai, Thomas Illig, Tim D. Spector, Jerzy Adamski, Nicole Soranzo, and Christian Gieger. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477(7362):54–60, August 2011.

[22] Johannes Kettunen, Taru Tukiainen, Antti-Pekka Sarin, Alfredo Ortega-Alonso, Emmi Tikkanen, Leo-Pekka Lyytikäinen, Antti J. Kangas, Pasi Soininen, Peter Würtz, Kaisa Silander, Danielle M. Dick, Richard J. Rose, Markku J. Savolainen, Jorma Viikari, Mika Kähönen, Terho Lehtimäki, Kirsi H. Pietiläinen, Michael Inouye, Mark I. McCarthy, Antti Jula, Johan Eriksson, Olli T. Raitakari, Veikko Salomaa, Jaakko Kaprio, Marjo-Riitta Järvelin, Leena Peltonen, Markus Perola, Nelson B. Freimer, Mika Ala-Korpela, Aarno Palotie, and Samuli Ripatti. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature Genetics*, 44(3):269–276, March 2012.

[23] Harmen H. M. Draisma, René Pool, Michael Kobl, Rick Jansen, Ann-Kristin Petersen, Anika A. M. Vaarhorst, Idil Yet, Toomas Haller, Ayşe Demirkan, Tõnu Esko, Gu Zhu, Stefan Böhringer, Marian Beekman, Jan Bert van Klinken, Werner Römisch-Margl, Cornelia Prehn, Jerzy Adamski, Anton J. M. de Craen, Elisabeth M. van Leeuwen, Najaf Amin, Harish Dharuri, Harm-Jan Westra, Lude Franke, Eco J. C. de Geus, Jouke Jan Hottenga, Gonneke Willemsen, Anjali K. Henders, Grant W. Montgomery, Dale R. Nyholt, John B. Whitfield, Brenda W. Penninx, Tim D. Spector, Andres Metspalu, P. Eline Slagboom, Ko Willems van Dijk, Peter A. C. 't Hoen, Konstantin Strauch, Nicholas G. Martin, Gert-Jan B. van Ommen, Thomas Illig, Jordana T. Bell, Massimo

Mangino, Karsten Suhre, Mark I. McCarthy, Christian Gieger, Aaron Isaacs, Cornelia M. van Duijn, and Dorret I. Boomsma. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nature Communications*, 6:7208, June 2015.

[24] Jaana A. Hartiala, W. H. Wilson Tang, Zeneng Wang, Amanda L. Crow, Alexandre F. R. Stewart, Robert Roberts, Ruth McPherson, Jeanette Erdmann, Christina Willenborg, Stanley L. Hazen, and Hooman Allayee. Genome-wide association study and targeted metabolomics identifies sex-specific association of CPS1 with coronary artery disease. *Nature Communications*, 7:10558, January 2016.

[25] Kirstin Mittelstrass, Janina S. Ried, Zhonghao Yu, Jan Krumsiek, Christian Gieger, Cornelia Prehn, Werner Roemisch-Margl, Alexey Polonikov, Annette Peters, Fabian J. Theis, Thomas Meitinger, Florian Kronenberg, Stephan Weidinger, Heinz Erich Wichmann, Karsten Suhre, Rui Wang-Sattler, Jerzy Adamski, and Thomas Illig. Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. *PLoS Genetics*, 7(8):e1002215, August 2011.

[26] Jan Krumsiek, Kirstin Mittelstrass, Kieu Trinh Do, Ferdinand Stückler, Janina Ried, Jerzy Adamski, Annette Peters, Thomas Illig, Florian Kronenberg, Nele Friedrich, Matthias Nauck, Maik Pietzner, Dennis O. Mook-Kanamori, Karsten Suhre, Christian Gieger, Harald Grallert, Fabian J. Theis, and Gabi Kastenmüller. Gender-specific pathway differences in the human serum metabolome. *Metabolomics*, 11(6):1815–1833, December 2015.

[27] Ayşe Demirkan, Peter Henneman, Aswin Verhoeven, Harish Dharuri, Najaf Amin, Jan Bert van Klinken, Lennart C. Karssen, Boukje de Vries, Axel Meissner, Sibel Göraler, Arn M. J. M. van den Maagdenberg, André M. Deelder, Peter A. C 't Hoen, Cornelia M. van Duijn, and Ko Willems van Dijk. Insight in Genome-Wide Association of Metabolite Quantitative Traits by Exome Sequence Analyses. *PLoS Genetics*, 11(1):e1004835, January 2015.

[28] Eugene P. Rhee, Qiong Yang, Bing Yu, Xuan Liu, Susan Cheng, Amy Deik, Kerry A. Pierce, Kevin Bullock, Jennifer E. Ho, Daniel Levy, Jose C. Florez, Sek Kathiresan, Martin G. Larson, Ramachandran S. Vasan, Clary B. Clish, Thomas J. Wang, Eric Boerwinkle, Christopher J. O'Donnell, and Robert E. Gerszten. An exome array study of the plasma metabolome. *Nature Communications*, 7:12360, July 2016.

[29] B. Yu, A. H. Li, G. A. Metcalf, D. M. Muzny, A. C. Morrison, S. White, T. H. Mosley, R. A. Gibbs, and E. Boerwinkle. Loss-of-function variants influence the human serum metabolome. *Science Advances*, 2(8):e1600800–e1600800, August 2016.

[30] Bing Yu, Yan Zheng, Danny Alexander, Alanna C. Morrison, Josef Coresh, and Eric Boerwinkle. Genetic Determinants Influencing Human Serum Metabolome among African Americans. *PLoS Genetics*, 10(3):e1004212, March 2014.

[31] Heli-Maria Salmi and others. Biochemical changes in inborn and acquired errors of metabolism. 2012.

[32] Sarah L. Spain and Jeffrey C. Barrett. Strategies for fine-mapping complex traits. *Human Molecular Genetics*, 24(R1):R111–R119, October 2015.

[33] Aharon Brodie, Johnathan Roy Azaria, and Yanay Ofran. How far from the SNP may the causative genes be? *Nucleic Acids Research*, 44(13):6046–6054, July 2016.

[34] So-Youn Shin, Eric B Fauman, Ann-Kristin Petersen, Jan Krumsiek, Rita Santos, Jie Huang, Matthias Arnold, Idil Erte, Vincenzo Forgetta, Tsun-Po Yang, Klaudia Walter, Cristina Menni, Lu Chen, Louella Vasquez, Ana M Valdes, Craig L Hyde, Vicky Wang, Daniel Ziemek, Phoebe Roberts, Li Xi, Elin Grundberg, Melanie Waldenberger, J Brent Richards, Robert P Mohney, Michael V Milburn, Sally L John, Jeff Trimmer, Fabian J Theis, John P Overington, Karsten Suhre, M Julia Brosnan, Christian Gieger, Gabi Kastenmüller, Tim D Spector, and Nicole Soranzo. An atlas of genetic influences on human blood metabolites. *Nature Genetics*, 46(6):543–550, May 2014.

[35] Seunggeung Lee, GonçaloR. Abecasis, Michael Boehnke, and Xihong Lin. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics*, 95(1):5–23, July 2014.

[36] Mary L. McHugh. The chi-square test of independence. *Biochemia Medica*, 23(2):143–149, 2013.

[37] Yana Sandlers. The future perspective: metabolomics in laboratory medicine for inborn errors of metabolism. *Translational Research*, 189:65–75, November 2017.

[38] Zhonghao Yu, Guangju Zhai, Paula Singmann, Ying He, Tao Xu, Cornelia Prehn, Werner Römisch-Margl, Eva Lattka, Christian Gieger, Nicole Soranzo, Joachim Heinrich, Marie Standl, Elisabeth Thiering, Kirstin Mittelstraß, Heinz-Erich Wichmann, Annette Peters, Karsten Suhre, Yixue Li, Jerzy Adamski, Tim D. Spector, Thomas Illig, and Rui Wang-Sattler. Human serum metabolic profiles are age dependent: Metabolic profiles associated with age. *Aging Cell*, 11(6):960–967, December 2012.

[39] Charu C. Aggarwal. *Outlier Analysis*. Springer New York, New York, NY, 2013. DOI: 10.1007/978-1-4614-6396-2.

[40] Ian J. Barnett, Seunggeun Lee, and Xihong Lin. Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies. *Genetic epidemiology*, 37(2):142, February 2013.

[41] E. Vartiainen, T. Laatikainen, M. Peltonen, A. Juolevi, S. Mannisto, J. Sundvall, P. Jousilahti, V. Salomaa, L. Valsta, and P. Puska. Thirty-five-year trends in

cardiovascular risk factors in Finland. *International Journal of Epidemiology*, 39(2):504–518, April 2010.

[42] Anne M. Evans, Corey D. DeHaven, Tom Barrett, Matt Mitchell, and Eric Milgram. Integrated, Nontargeted Ultrahigh Performance Liquid Chromatography/Electrospray Ionization Tandem Mass Spectrometry Platform for the Identification and Relative Quantification of the Small-Molecule Complement of Biological Systems. *Analytical Chemistry*, 81(16):6656–6667, August 2009.

[43] S. Goto, H. Bono, H. Ogata, W. Fujibuchi, T. Nishioka, K. Sato, and M. Kanehisa. Organizing and computing metabolic pathway data in terms of binary relations. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 175–186, 1997.

[44] LinS. Chen, Li Hsu, EricR. Gamazon, NancyJ. Cox, and DanL. Nicolae. An Exponential Combination Procedure for Set-Based Association Tests in Sequencing Studies. *American Journal of Human Genetics*, 91(6):977–986, December 2012.

[45] Nenad Blau. Genetics of Phenylketonuria: Then and Now. *Human Mutation*, 37(6):508–515, June 2016.

[46] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387, October 2012.

[47] Emil Julius Gumbel. *Statistics of Extremes*. Courier Corporation, July 2004. Google-Books-ID: kXCg8B5xSUwC.

[48] Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J. Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology*, 5(1):21, January 2011.

[49] Peter H. WESTFALL. Kurtosis as Peakedness, 1905 – 2014. R.I.P. *The American statistician*, 68(3):191–195, 2014.

[50] H. Frederik Nijhout, Janet Best, and Michael C. Reed. Escape from homeostasis. *Mathematical Biosciences*, 257:104–110, November 2014.

[51] Tanja Cvitanović, Matthias C. Reichert, Miha Moškon, Miha Mraz, Frank Lammert, and Damjana Rozman. Large-scale computational models of liver metabolism: How far from the clinics? *Hepatology*, 66(4):1323–1334, 2017.

Glossary

acidosis A metabolic state whereby the blood and tissue are acidic (.i.e, pH is less than 7) for extended periods of time. 20

allele 12, 13, 16, 18, 22, 36, *see* variant

autosomal 3, 30, 32, 34, *see* autosome

chromosome A contiguous strand of DNA. In humans there are 23 chromosome types and each individual inherits a version from both parents. Thus, the total number of chromosomes posessed by an individual is 46. 3, 10, 11, 13, 30, 32

codon A three-nucleotide long unit of information within a protein coding genomic region. Each codon translates to an amino acid residue present in the encoded protein or the beginning or end of the coding region. 7, 15, *see also* transcription

dimorphism A binary difference i.e sexual dimorphism - difference between sexes. 19

diploid Diploid organism possess two sets of distinct chromosomes. That is, each chromosome is present in two instances. 13

enzyme A protein which catalyzes chemical reactions. 16, 46

exome The total set of exon sequences. 1, 19, *see also* exon

exon A subsequence of the gene which is included in the final transcript RNA. 7, 9, 15, 22

gene A gene is a region encoding the amino acid sequence of a specific protein. Typically one gene encodes one protein. 1, 20–22, 30, 32, 34–36, 39–44, 49

genome-wide association study A study where variation across the complete genome is examined with respect the phenotype of interest. i, 15, 18, 19

haploid Possesing a single copy of each chromosome. 11, 13, *see also* diploid

heritability The amount of variance in the trait explained by common descent. 15

heterozygosity *see* heterozygote

heterozygote A term used to denote a diploid individual, who possesses two different alleles within the same locus. That is, whose paternally inherited allele differs from the maternally inherited. 10, *see* diploid, &

heterozygotic *see* heterozygote

homologous recombination Homologous recombination is the molecular mechanism of chromosomal recombination in meiosis. 13, *see also* meiosis

homozygosity 13, *see* homozygote

homozygote A term used to denote a diploid individual, who possesses two identical alleles within the same locus. That is, whose paternally inherited allele is identical to the maternally inherited. *see also* diploid, &

homozygotic *see* homozygote

intron A subsequence of the gene which is excluded in the final transcript RNA. 7, 9, 15

ketosis A metabolic state whereby a significant fraction of the body's energy is produced by consuming compounds called ketone bodies. These ketone bodies are produced during fasting or other conditions of carbohydrate depletion. 20

linkage analysis Linkage analysis is an genetic association study design, which associates traits to genotypes by analyzing pedigrees.The likely genomic location of causal variants can be inferred using the observed distribution of traits across the pedigree and the mixing properties of meiosis. 1, *see also* meiosis & homologous recombination

locus The region where the variant site lies. This region can be for example a gene or a subregion of a gene. 13, 16

meiosis The process of gamete (reproductive cell) generation. The process can be summarily described as non-conservative cell division process of a diploid cell resulting in haploid daughter cells. The ancestral chromosomes of the parent cell undergo pairwise recombination (homologous recombination) and finally duplication. 13

Mendelian Traits which inheritance follows the Mendelian inheritance model. 1, 13, 17, *see* Mendelian inheritance

metabolite A small-molecule compound (eg an amino acid,lipid,peptide). 1, 16, 18–23, 29, 30, 34, 39, 40, 43–46

metabolomic 16, 18, *see* metabolomics

mGWAS GWAS, where phenotypes are small-molecule concentration in a biofluid or tissue. 19, 21, 22, 30, 35, 36, 38–42, 44, 46, 49

missense mutation A point mutation which results in a change of protein amino acid sequence. 15, 44, *see also* codon

mutation An event whereby a genetic sequence is altered. 15, *see also*

non-sense mutation A point mutation resulting in a premature stop-codon. see 15, 44

non-synonymous mutation A mutation leading to an change in amino acid residue or a premature stop-codon. 22

pathway A chain of interacting molecules organizing the biological function of the cell. For example, signalling pathways consist of protein molecules interacting pairwisely to relay information across the cell to initiate process. Or metabolic pathways which consist of sequence of chemical reactions resulting in a desired product. 3, 20, 21, 44, 46

phenotype A set of individual traits which are defined or influenced by the genotype. 10, 15, 16, 18, 22, 43, 44

silent mutation A mutation substituting a codon with it's synonym, thus leading to no change in protein amino acid sequence. 15

single nucleotide polymorphism Single Nucleotide polymorphism; An alteration of single base pair in the DNA sequence. 16, 21, 44

splice-site mutation Splice-site mutations refer to variants which confound the intron-exon boundary leading to inappropriate inclusion of introns or exclusion of exons to the transcript. 15, *see* splicing

splicing Splicing refers to the process of removing introns from the raw transcript to form the final transcript. 7, 15

transcriptomic *see* transcriptomics

transcriptomics The study of gene expression activity in vivo. The amount of transcription is measured per gene and treated as measure of activity for the gene. Then, depending on the experiment, differences in activity are analyzed with respect to other variables of interest. *see also* transcription

variant A specific version of a gene or a genomic region Though the term can be used to refer to multiple types of differences (eg deletions or insertions), in the context of this work only SNPs are considered. 1–3, 13, 15, 17, 19–23, 30, 32, 35–44, 49, 50

Acronyms

CAD coronary artery disease 13, 19

CL chromosomal linkage 13, *see glossary:cl*

CVD cardiovascular disease 27

DNA deoxyribonucleic acid 3, 7

EPS extreme phenotype sampling 22

EXAC Exome aggregation consortium 30

GGM gaussian graphical machine 19

GWAS genome-wide association study 1, 16, *see glossary:* genome-wide association study

H-NMR hydrogen nuclear magnetic resonance 16, 18

IEM inborn error of metabolism 1, 3, 17, 20–23, 27, 30, 32, 34–37, 39–41, 44, 46, *see glossary:* inborn error of metabolism

LD linkage disequilibrium 12, 13, 21, *see glossary:ld*

LOF loss-of-function 1, 15, 17, 20, 22, 43, 44

MAF minor allele frequency 22, 36, *see glossary:* minor allele frequency

mGWAS metabolic genome-wide association study 3, *see glossary:* mGWAS

mQTL metabolic quantitative locus 18

mRNA messenger RNA 7, *see*

MS mass spectrometry 16, 18

QTL quantitative trait locus analysis 16

RNA ribonucleic acid 7

SNP single nucleotide polymorphism 15, 21, 35, *see glossary:* single nucleotide polymorphism

T2D type 2 diabetes 13